A Bottom-Up Approach to Linguistic Persuasion in Advertising:
Modeling the effects of TV ads using automated text analysis

Nick Beauchamp*
NYU Department of Politics

9/9/11

Abstract

This paper presents a new, bottom-up approach to understanding how the linguistic contents of television advertisements determine their persuasive effects. To deal with the huge variety of political advertising, existing econometric approaches generally require first reducing this data via factor analysis, scaling, or expert categorization. However, these reductions both limit the generality of any results, and often fail to find meaningful effects, particularly when there are more than a few underlying dimensions. As an alternative to such data reduction, I develop a simple but effective one-at-a-time regression technique to estimate the effect of each unique TV ad on survey-measured vote intention during the 2004 US presidential campaign. I find that these effects were in aggregate substantively significant, particularly on the Democratic side, and would have been substantially larger had the campaigns better selected which ads to broadcast. To understand why some ads were more effective, I develop a number of new automated text analysis procedures to scale the ads’ words according to their persuasive effects. This scaling summarizes the textual characteristics of the more effective Democratic and Republican ads and shows that the two sides used very different persuasive strategies. To validate this approach, I develop related procedures to predict the effects on vote intention of new, unmeasured ads based only on their textual similarity to field-tested ads. This constitutes both a powerful campaign tool, and a demonstration that automated text procedures can, if properly designed, provide a wide-ranging method for inferring the effects of the ever-varying linguistic events common in persuasion.

*Email: nick.beauchamp@nyu.edu. Web: nickbeauchamp.com. The author would like to thank for their generous and helpful comments on this paper or a presentation of it: Jonathan Nagler, Michael Laver, Nathaniel Beck, Arthur Spirling, Gary King, Burt Monroe, Adam Bonica, Alex Herzog, and the participants at the Saint Louis Area Methods Meeting.
1 Introduction

How do voters make their choices? The answers to this question are nearly as numerous as the social scientists working on it. On one side we have affective theories, where voters make their choices based on emotional, often non-conscious feelings, and decision changes are effected by media imagery, music, and narratives (Brader 2006, Mutz 1996, Mutz & Reeves 2005, Mutz 2007). On the other side we have rational, information-based decision making, where citizens update their views in a Bayesian fashion based on new data (Lupia & McCubbins 2000, Ferejohn & Kuklinski 1990, Sniderman, Brody & Tetlock 1993, Page & Shapiro 1992, Popkin 1991). And in between, we might position the theories of persuasion more properly: agenda-setting, priming, framing,\(^1\) and (edging towards the more rational end) biased or strong-prior Bayesian updating.\(^2\) Most research has tested these theories separately, either in the laboratory or via survey data, but it is likely that all of these effects are present in various degrees during a major election campaign. How are we to determine which strategies are dominant, and which cause voters to diverge from the baseline established by their party ID and demographic background? How can we build a general theory of persuasion (taken as an umbrella term) that is sensitive to the combined and competing means of changing political opinion?

A bottom-up approach to explaining persuasion using automated textual analysis offers a new tool for answering these questions, allowing us to work up from words to themes and strategies in a general, systematic, and statistically well-grounded way. The result, as I show below, is that

---


\(^2\)For instance, Geer (2006) suggests that negative ads are more useful than their reputation suggests, because they are actually quite informative for low information voters, due in part to the fact that negative claims require more substantiation to be palatable.
presidential advertising campaigns successfully employ a mixture of very different strategies, along with many more unsuccessful ones: In 2004 the Kerry campaign (and Democrats more generally) engaged in many different persuasive efforts, but the most successful was a policy-oriented framing strategy, where the issue of health care was emphasized and information about prescriptions was promulgated. Bush and his supporters, on the other hand, most successfully countered this not with alternative framing efforts, but with a more affective, partisan strategy, emphasizing the weak “Democrat” brand, making the infamous personal “swift boat” attacks, and gaining benefit from the libertarian candidate peeling disaffected voters away from Kerry. All of these strategies, and many more, could be identified by qualitative analysis of the advertisements, but only the bottom-up, textual approach allows us to distinguish the successful ones from the unsuccessful, and to recognize that the true combat (which largely favored Kerry, as we will see) was between asymmetric strategies: framing and information on the one hand, affect and negativity on the other.

The approach here uses survey and advertising broadcast data, which ensures both good external validity, and a wide array of real advertising strategies to explore. But because there are hundreds of unique ads broadcast during a given political campaign, most of which are shown for only a few days in a few markets, detecting the effect of each unique ad has usually been deemed infeasible. Instead, the researcher usually begins with a hypothesis, categorizes those ads according to that hypothesis (negative/positive ads, e.g.), and then employs traditional cross-sectional and/or time-series analysis to those categorical variables. But this approach is less than ideal: first,

---

3 Although the experimental approach has gained increasing popularity as a means of testing theories of persuasion, it has a number of drawbacks. Iyengar and others (e.g., Iyengar, Peters & Kinder (1982), Iyengar & Kinder (1987)) have done a solid job establishing the effects of priming and framing in the laboratory, but there remain deep issues of external validity in generalizing from attentive college students to inattentive TV watchers. And even with the improved validity of field experiments, generalizing from a single well-varied ad to the vast panoply of real treatments is a daunting task.

4 Although they do not all focus on ads, a wide variety of analyses of mass political persuasion rely on simple categories. Most rely on expert categorization, which in turn appears to depend largely on keywords and simple themes. See, for instance, Jacobs & Shapiro (2000), Popkin (1991), Arnold (2006), Geer (2006), Hillygus & Shields
because it limits the researcher to existing hypotheses or encourages data mining; second, measured effects may be attenuated by mixing in ineffective ads with the (unknown) effective ones; third, even expert categorization presumably adds significant measurement error; and fourth, even apart from error, these categories may be applied in a highly subjective fashion: one person’s negative ad may be another’s call to action.

The approach developed here works the opposite direction: not top-down from hypothesis to categorization to measurement, but bottom-up, from measuring individual ad effects, to generalization based on those measurements and the textual content of the ads. Such an approach presents difficulties, though: with only a few ads, measuring each of their effects might be feasible, but generalization would be difficult; whereas measuring the effects of many ads is not possible with existing methods. Thus a new technique is developed to allow us to measure the effect of each of hundreds of unique ads: one-at-a-time regression. The challenge, of course, lies not in asserting these ad effects, but in the statistical demonstration of their validity. To this end, both simulation and out-of-sample testing is employed to show that one-at-a-time regression works better than the alternatives in predicting vote intention based on ad broadcast data. And unlike many other approaches, it also provides an estimated effect for every single unique ad.

But even having determined the aggregate accuracy of these effect measurements, the problem of generalization returns: what accounts for the variation in ad effects? As before, the researcher could comb through the most effective ads, looking for any discernible patterns, but this would lead to hypotheses that were ad hoc, difficult to test, and limited in scope. Instead, another set of new techniques are devised, using Bayesian and machine learning methods. These techniques allow

us to estimate the most effective words from the set of ad effects, but as before, it is essential to establish that these estimates are empirically sound. So as before, the models are validated through out-of-sample prediction: first by showing that word frequencies can predict abstract categories like “negative” ads, and second, by showing that the persuasive effects of ads can be predicted from their textual similarity to ads with known effects. After this validation – an important methodological development in its own right – a number of different word scalings are developed and employed to infer the characteristics of successful ads. These word-based inferences show the differing Democratic and Republican strategies, persuasive effects that would be difficult or impossible to have detected with traditional, top-down tests.

2 Data

The data consist of two parts: first, the Campaign Media Analysis Group (CMAG) has collected broadcast information on every TV ad run during the 2004 presidential campaign,\(^5\) down to the time and location of each broadcast. CMAG also provides “storyboards” with transcripts of each unique ad, totalling 359 for the time period examined here. Finally, CMAG has itself categorized these ads according to many topics relevant to various theories of persuasion: whether the ad is negative, issue-oriented, mentions women or health, etc. The second part of the data comes from the 2004 Annenberg survey, which polled nationally on a rolling basis during the entire campaign period.

The dependent variable here is vote intention; in addition, a variety of control variables are

\(^5\)Which, for the present purposes, is taken to run from mid-July, when advertising begins to ramp up, through election day in November.
also used. Since the key independent variable, advertising, varies by week and by media market, the data are aggregated at the week-region level. Thus for each observation, the DV is the survey-measured intended Kerry vote as a proportion of the intended two-party vote in that week-region, while each control is the mean survey response (e.g., average number of women) for that week-region. Most importantly, the 359 ad count variables are the total number of broadcasts for each unique ad in that region during that week. Although the Annenberg survey is large, even at this level of aggregation there are only about 50 survey responses per aggregated observation, which means that although the data are clearly time-series cross-sectional, including the temporal aspect of the data does not increase explanatory power. Each observation is sufficiently noisy that there is no autocorrelation between region observations from one period to the next, nor is there autocorrelation between residuals from a pooled model regressing vote intention on controls, nor do the estimates from the pooled model differ from those using standard time-series models. Thus, at least at this level of aggregation, the data can be treated as a single pool, a fact which will be put to good use when doing out-of-sample predictions to test model accuracies (although there will be one forward-in-time out-of-sample test later on).

---

6Specifically, party ID, ideology, age, employment, race, religiosity, south, and urbanism.

7The markets are Nielsen Designated Market Areas (DMA). Most ads tend to only be run for a week or so, many even beginning at the start of a week, making that an appropriate time unit.

8It is possible to have weighted broadcasts by estimated cost or by survey results on TV watching rates. For instance, Prior (2007) or Huber & Arceneaux (2007) at times weigh by population, cost, TV program choice, or attention to advertising. However, they acknowledge that these weighing may add an extra layer of noise, and for the present purposes less noise is preferable to slightly more accurate measures of attention.

9Had there been temporal correlations, the data could also have been detrended in various ways to allowed pooled testing, but this was not necessary. In addition, fixed effects for regions are not included (as dummies or in a TSCS model). This is for a number of reasons: First, with 50+ regions and only 16 time units, the fixed effects will absorb most of the variance we might try to explain. Second, summary statistics of the dependent and control variables for the most part do not differ significantly by region unit. And third and most importantly, including the fixed effects worsens out-of-sample prediction of the DV (using the procedures described below), showing that fundamentally, they do not improve model fit.

10Treating the data as pooled leads to greater problems of endogeneity, but since we are interested in per-ad effects, reverse-causation arguments are somewhat implausible: Do we expect that Kerry’s team sees they are underperforming in some regions and then deploys a health-related ad only in those regions? This is unlikely; why not deploy an effective ad everywhere. The sheer number of different ads suggests that campaigns have complex and, as will be shown, quite often erroneous notions of what will be effective where. Another argument against endogeneity is that,
3 The problem of many variables

Because there are 359 unique ads and only 859 week-region observations, estimating individual ad effects via traditional multiple regression would lead to extreme over-fitting. That is, we might get a high $R^2$ and many significant effect coefficients on our ad variables, but were we to use such coefficients to predict out-of-sample, the prediction would be terrible (as we will see). There are a number of established approaches to dealing with this problem, but none of them are adequate to the present purposes. The most common approach is to reduce the number of variables by first categorizing them, as explained above; these category count-variables could then be included in a single regression, or added stepwise, or only those relevant to a specific research hypothesis could be included. But even if this produced more manageable data (which, as we will see, it does not) it is vulnerable to cherry-picking or data-mining, and limits the generality of what can be discovered.

A more systematic response to the problem of numerous variables is dimensional reduction via an eigenvector method. Although there are many (most of which produce similar results), the most common is probably principal component analysis (PCA), which takes the entire 359-dimensional space of ad variables and reduces it to a few dimensions corresponding to the highest eigenvalue eigenvectors of the original variable matrix. But these derived variables lead to problems with generalization and interpretation, particularly when there are many initial dimensions, and it generally works only when the data are inherently low-dimensional, which is not the case here.

The major alternative to variable reduction via eigenvector decomposition is simply variable
selection. Again there are a host of possible algorithmic approaches to this familiar problem, but perhaps the most popular new method is LASSO (least absolute shrinkage and selection operator), which shrinks most of the estimated coefficients to zero, leaving a small subset of purportedly significant variables (Tibshirani 1996). It begins with the usual OLS coefficients, but seeks to minimize the following:

$$\arg \min_\beta \|Y - X\beta\| - \lambda \sum_{k=1}^{j} |\beta_k|$$

That is, the total size of the beta coefficients is constrained, which leads to many of the insignificant ones being reduced to zero. The approach is good if indeed most of the independent variables have no direct effect on the dependent variable, but works less well if, as appears to be the case here, many or most of the variables have small but not nonexistent causal contributions. Furthermore, generalization is once again a problem: the result would be a few ads with significant effects and most with none, which would limit us to speculating about what traits underlie these particular successes.

As an alternative to these existing reductive approaches, I propose a new method, one-at-a-time regression. It allows us estimate the effect every ad, a characteristic which will be essential text-based inferences later on. But in addition, it simply works better: both simulations and out-of-sample tests with the ad data show that, in cases like these where the data are noisy and the variables many, one-at-a-time regression can be superior to multiple regression as well as variable-reduction techniques like categorization, PCA, and LASSO.

The procedure is straightforward: rather than estimate $y = X\beta$ on all $K$ independent variables

---

11Stepwise variable selection, ridge regression, least angle regression, etc. As with eigenvector methods, these procedures all generally produce similar results, usually eliminating all but the highest-significance coefficients.
at once, one instead makes $K$ independent OLS estimates:

$$y = \beta_k X_k + Z\Gamma_k$$

(1)

where $X_k$ is an ad count variable, and $Z$ are the control variables, if they exist. The procedure is inspired by the recent success of ensemble methods (Dietterich 2000, Dietterich 1997, Schapire 2003, Sebastiani 2002, Page 2008), where numerous studies have now demonstrated the ability of ensembles of different models to out-perform single, more complex models, often because the former can capture diverse characteristics of the data without over-fitting.\(^{12}\)

Thus the one-at-a-time approach can be considered the simplest possible ensemble of models, an idea drawn in part from Page (2008).\(^{13}\) Obviously the major concern here is omitted variable bias, which is discussed in Appendix A. The short answer is that the proof lies in the out-of-sample testing (both simulated and real), which shows that whatever the drawbacks, in these cases one-at-a-time is superior. It appears that when the variables are so numerous, the spurious (over-fit) inter-variable correlations are more damaging than the omitted variable bias. As Appendix A shows, although one can somewhat correct for omitted variable bias when one knows the variables being omitted, this merely drags results back down to the level of an over-fitted multiple regression.

Precisely because of worries like omitted variable bias, we cannot simply trust the significance levels of the individual coefficients to establish the accuracy of the model. The best general-purpose

\(^{12}\)Although even this can be pushed too far. Random forests go beyond even the one-a-time procedure, creating many more decision parameters than there are independent variables, and allowing sub-models to vote on overall decisions. But even this very popular approach has run into problems with over-fitting with real-world data (Breiman 2001, Segal 2004).

\(^{13}\)Another basis for this approach is the Spirtes algorithm (Spirtes & Glymour 1991), a method drawn from the causal inference literature and based on the work of Pearl (Pearl 2000, Pearl & Verma 1991, Pearl 1995). However, while the Spirtes algorithm investigates triplets of variables to eliminate most causal linkages, the one-at-a-time procedure examines pairwise, and posits that it is best to retain every possible causal connection, no matter how weak it appears.
way of comparing these various models is via out-of-sample testing: fit the model on a subset of the
data, and predict the dependent variable for the rest. For most of these approaches the prediction
step is straightforward, but for one-at-a-time-regression, one needs a method for producing a single
\( \hat{y} \) out of these \( K+1 \) regressions. Based again on the successes of ensemble methods, the approach
here is via simple averaging:

\[
\hat{y} \equiv \bar{\hat{y}}_k = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_k X_k + Z \hat{\Gamma}
\]  

That is, the predicted \( \hat{y} \) is simply the average of each predicted \( \hat{y}_k \) for each of the \( K \) one-at-a-time
regressions, plus \( Z \hat{\Gamma} \), where \( \hat{\Gamma} \) is vector of coefficients from regressing \( y \) on the controls \( Z \) alone.

The success – and limitations – of the one-at-a-time method can be seen via Monte Carlo
simulation. Figure 1 shows the out-of-sample predictive accuracy (and thus the accuracy of the
inferred coefficients) for four methods: multiple regression, LASSO, one-a-time regression, and the
simple (in-sample) mean \( \bar{y} \). The true model is \( y = X \beta + \epsilon \),\(^{14}\) and the metric is the root mean
square error (RMSE) \( \sqrt{1/N \sum (y_i - \hat{y}_i)^2} \), as it will be throughout this paper. As the noise level
\( \epsilon \) is gradually increased in generating \( y \), the ability of each of the models to predict \( y \) goes down.

Figure 1 shows the RMSE minus the noise level\(^ {15}\) as the noise level increases; lower scores are
better, and the model with the lowest score for each noise level is the one that does the best job
inferring \( \hat{y} \), and therefore the true coefficients \( \beta \). The results are interesting: for the lowest levels
of noise, multiple regression works best, reassuringly; for higher noise levels, LASSO does the best

\(^{14}\)Specifically, in this case there are 10 independent variables, which are lightly correlated with each other and
five of the coefficients are set to 0, while the other five are drawn from a uniform distribution for each run. The
intercorrelation should give multiple regression a leg up on the one-at-a-time method, while setting five coefficients
to 0 should give LASSO a leg up. However, the results are much the same if the variables are not correlated; if the
coefficients are all drawn from a uniform distributions; and if there are more or fewer than 10.

\(^{15}\)That is, for each stage (each of which is iterated 10,000 times), \( \epsilon \) is drawn from a uniform distribution between
\(-m \) and \( m \); for each stage, \( m \) increases by one. Figure 1 shows the level \( m \) on the x axis, and the RMSE minus \( m \)
on the y axis.
job; but for yet higher noise, one-at-a-time regression does the best, across a large range of noise levels. At the highest noise level, none of the methods are able to recover any information about the beta coefficients, and one is best off eschewing a model and simply guessing \( y \) using its in-sample mean. Given the popularity of LASSO for high-variable, high-noise situations, these simulations show that one-at-a-time regression has an important role where the noise and variables are quite high, as long as one is willing (as here) to have coefficients for every single variable.

[Figure 1 about here.]

Of course, simulations cannot tell us which region of this graph empirical data occupy, and thus which method is best employed. To establish that, a variety of methods are again tested via out-of-sample prediction using the ads data. These data are treated as pooled, so for each run, the in-sample is a randomly selected 95% of the 859 observations, while the out-of-sample testing group is the remaining 5%.\(^{16}\) As before, the models are fit in-sample, and then the dependent variable, vote intention, is predicted for the out-of-sample set and compared via RMSE to the true vote intention. Since results fluctuate with each sample, the best test of the success of an ads-including model is whether that model, with the ads estimates (however they are made), produces a more accurate prediction of vote intention than does the baseline model using only the controls. Given that the controls-only model already accounts for more than 60% of the variance, this is a high bar to set, and it is easy for a model (such as multiple regression) to appear to have generated significant coefficients on ad variables, but fail to predict \( \hat{y} \) any better than the controls, due to over-fitting.

\(^{16}\)This approach works better than leave-one-out cross-validation, since it produces smoother estimates without the fluctuations due to single outliers; it is also appears less vulnerable to over-fitting.
The results of these out-of-sample tests can be seen in Figure 2. Each line shows the percent of out-of-sample runs (out of 1000) where the specified model produces a lower RMSE than the controls-only model. By binomial chance, we would expect a model adding a small random perturbation to fall between 46 and 54 95% of the time; anything outside of these bounds is likely to have out- (or under-)performed due to being genuinely better (or worse) than the baseline. To give a sense of the magnitude of these effects we can remove a control variable: Party ID is of course the most significant one, and row 2 shows that without that variable the model does worse on 100% of the out-of-sample runs. On the other hand, removing any one of the other controls doesn’t irreparably damage the model; on average, removing a non-Party-ID control means the new model still beats the baseline 40% of the time – significantly worse, but not as bad as Party ID.

But the important question is what effect the ads have, and which models are able to detect these effects. One might try adding a number of category variables into a standard the regression (such as “negative Democratic-sponsored” or “women-related Republican-sponsored”) to explore the effects of these various persuasive strategies. But row 4 shows that adding all the CMAG category variables significantly worsens the model, demonstrating no discernible ad effect. This also holds for most subsets of the CMAG variables (not shown in Figure 2), although one can of course find a few unrelated ones that will work together if one cherry-picks long enough. And adding all the raw count variables for the ads in a single multiple regression works terribly of course – with 359 ad variables alone, the model severely over-fits on the in-sample data.

[Figure 2 about here.]

\footnote{A category variable counts the total number of broadcasts of ads of that category in that region during that week. CMAG categories include topics such as: whether the ad is an attack or self-promotion; whether it cites evidence, promotes action, is a rebuttal, includes images of the candidate, is policy or personal; and issues such as the economy, women, education, health, the poor, the elderly, crime, morals, or drugs. For estimation, each category variable must be spit into two based on the sponsoring party, since we expect them to have opposite effects on vote intention.}
Facing these problems with over-fitting, the standard solution is variable reduction. If we take the principal components of the ad count variable matrix and add the top few components as variables, we do in fact do slightly better than the baseline (row 6 of Figure 2 shows 6 components, but results are only slightly worse for fewer). But only barely, and we are still left with problems of interpretability. If we instead use the LASSO method, we do better than the baseline, but not quite significantly so. And again, even though in this case we know which ad variables are included (usually the ones with the highest standardized coefficients, roughly), we again have problems generalizing from that small set to any broader theory of persuasion.

Finally, if we employ the one-at-a-time method, averaging the predictions from each of the individual ad regressions, we do far better than the baseline, and much better than any of the other methods. The one-at-a-time method is vastly superior to the other approaches at estimating ad effects and making accurate out-of-sample predictions. This approach also importantly produces effect sizes for each ad – although of course each individual coefficient estimate is quite noisy. (The remainder of Figure 2 will be returned to shortly.)

The out-of-sample testing has established that there is a statistical relationship between ads and vote intention, if measured correctly. We must now establish the substantive magnitude of the aggregate ad effect, and account for the variation in individual ad effects. One obvious approach would be simply to list the most Democratic- and Republican-persuading ads and attempt to generalize from them. This immediately raises another measurement issue, though: the highest (absolute value) coefficient ads will include ads that are both truly persuasive, and those that have weak real effects, but happened to have high errors in measurement. One way to mitigate this prob-

---

18 A simple means-difference test for the 1000 RMSE’s from each method finds significant differences between all of them, although this approach is a bit less reliable inasmuch as one could always run more and more out-of-sample tests until even the smallest accidental difference due to quirks in the data became significant.
lem is through shrinkage: as in a multilevel model, we assume that the true ad effects are normally distributed around some mean, and each is then measured with a (normally distributed) error; we then shrink the estimated coefficients back towards the mean in proportion to the estimated measurement error. The shrinkage employed here is Bayesian: if we assume the true coefficients are distributed $\beta \sim \mathcal{N}(\bar{\beta}, \tau^2)$ and each measured coefficient is drawn from $\hat{\beta}_k \sim \mathcal{N}(\beta_k, \sigma^2_k)$, then the shrunk coefficients are

$$\hat{\beta}'_k = \frac{\sigma^2_k \bar{\beta} + \tau^2 \beta_k}{\sigma^2_k + \tau^2} \approx \frac{\hat{\sigma}^2_k \bar{\beta} + \tau^2 \hat{\beta}_k}{\hat{\sigma}^2_k + \tau^2}$$

where the latter quantity contains the estimated values: $\bar{\beta}$ is the mean of the measured coefficients and $\hat{\sigma}^2_k$ is the OLS-derived variance on the $k$th coefficient. The only unknown quantity is $\tau^2$, the true variance, which is estimated by cross-validation. The practical result of this is that ads with apparently high effects, but also high variance (usually due to having been run only a few times) are shrunk back, leaving something at least closer to the real effects. As can be seen on row 9 of Figure 2, the shrunk coefficients predict out-of-sample slightly better than the unshrunk coefficients.

Now that we have estimated the ad effects as well as possible, we can return to the question of explaining them. Knowing which are the most effective ads is good, but why are some ads more effective than others? The best approach, as we will see, will involve estimating the highest-effect words and inferring strategies from those terms. Simply listing the top Democratic- and Republican-effect ads doesn’t work: many Republican-sponsored ads are among those with high Democratic effects, and vice versa. This is presumably due to some combination of backfiring results and still-noisy measurements of individual ad coefficients. So what can we say more broadly about the

\footnote{Again, one could also have done this in a single multilevel step, but the results would be similar and, inasmuch as they differ, one should prefer the cross-validated optimum to the MLE optimum.}
differing effects of Democratic- and Republican-sponsored ads? Surprisingly, this difference is very small. Figure 3 shows a histogram of ad effect sizes by party, and as can be seen, the vast majority of ads have very small effects, and many Democratic-sponsored ads have Republican-helping effects, and vice versa. A t-test on these distributions finds that the two categories in fact have means that are not significantly different, although a more discerning non-parametric Kolmogorov-Smirnov test shows that they are in fact slightly different. However, the difference is tiny: most ads have little effect, and many have opposite their intended effect.

[Figure 3 about here.]

Clearly the campaigns are not nearly as effective as they might believe. But just because many of their attempts backfire does not mean that the ads in aggregate do not have large effects. Table 1 shows the estimated effect on vote intention of shifting each variable one standard deviation or from its 30th to 70th percentile. For the ads, a single aggregate variable is constructed as \[ v = \frac{1}{j} \sum_{j} \hat{\beta}_k X_k, \] ie, the mean prediction (full-sample, in this case). While all the other control variables (apart from Party ID) shift the outcome by 1-2 points (which means the difference between the candidates would increase by twice that, 2-4 points), the unified ad variable is twice as “effective”: a uniform overall increase in ad spending changes vote intention by 2-4 points – a large amount in a tight race.

[Table 1 about here.]

Another way to see the significance of these ad effects is to consider a hypothetical case: say a candidate had the luxury of going back and doing her campaign over, using the same ads,
but selecting only the better ones based on what she learned here. How much could she improve her outcome? Figure 4 shows the estimated outcome were one candidate to buy only the most effective ads, those above some cutoff. Were either candidate to have bought no ads at all, the outcome would be virtually unchanged, since both sets of ads have nearly 0 mean effect (not shown in figure). But had the Kerry campaign bought exactly as many ads, but only those with a coefficient greater than 0 (ie, only those estimated to have a pro-Democratic effect), his outcome would have been significantly, if slightly, improved. Were he to have bought only the top 25% best performers (coefficients > 0.001), he would have had a significant chance of winning the election. And were he to have broadcast only the top 1%, the outcome would have been dominantly in his favor. Interestingly, although the pattern is similar for Bush, the effects are much weaker, with only a slight improvement even from choosing the highest-performing ads. Thus, although both campaigns were about equally ineffectual, the pro-Democratic ads were much more effective, had only they known better which to choose.

21

[Figure 4 about here.]

But this simulated strategy in reality does neither the researcher nor the campaign much good. Elections cannot be run again, and even if they could, that still would not let us explain why some ads are more effective than others. One direct approach would be to look at mean differences in effect, this time not between Democratic and Republican-sponsored ads, but between more specific and substantively interesting categories: negative Democratic-sponsored ads versus

20 That is, the candidate runs the exact same number of broadcasts, only distributed differently among their ads. 21 Presumably these strategic outcomes are still exaggerated by the fact that many of the largest estimated ad effects are still due, in part, to measurement error on top of their true effect size. Shrinkage was designed to mitigate this, but of course it cannot have eliminated it, so these estimates due to strategic buying should perhaps be taken as upper bounds.
positive Democratic-sponsored ads; Republican ads emphasizing women or the war, versus those that don’t, etc. But given how poorly the CMAG category variables performed in the out-of-sample testing, we shouldn’t be surprised to find that these means-difference tests fail to find significant differences in almost all cases. Instead, for this bottom-up approach to provide any general explanations, we must turn to a statistical analysis of the content of those ads, their text.

4 Inferring causation from text

To explain the varying effects of ads using their textual contents, the task is two-fold: first we must establish a statistical relationship between their text and their effects, and only after this is done can we be confident that these methods allow us to infer the words that characterize the most effective ads and strategies. Automated text analysis generally takes one of two forms, broadly speaking: categorization, or scaling. The former clusters documents into groups, and can also provide the words that are most characteristic of each group. The latter begins with the assumption that documents and words can be placed in a single or multidimensional space, such as a left-right ideology, and employs various techniques to best fit the document data to such a space. Text analysis may also be supervised or not: the former begins with categories or positions already assigned to some documents (by experts, say) and uses that information to categorize or scale the rest, while the latter assumes no such knowledge. But supervised or not, categories or scales, these automated text procedures share a weakness when it comes to traditional causal, empirical models.

The unsupervised approaches are good for exploratory work, but the interpretation of their output

\[\text{For unsupervised categorization in political science, see Grimmer & King (2011) and Grimmer (2010). For unsupervised scaling, see Monroe & Maeda (2004) and Slapin & Proksch (2008). For supervised scaling, see Laver, Benoit & Garry (2003) and Beauchamp (2011).}\]
rests entirely with the researcher – simply replacing traditional qualitative analysis of the texts with qualitative analysis of an automated algorithm. If this output is to be employed in an econometric model, the researcher must accept it as is, and do her best to justify whatever interpretation she is making about the meaning of those numbers. Supervised approaches would seem to have an advantage, but rather than supervise the output to be most useful for a substantive model, researchers – particularly in computer science where these techniques were begun – generally have contented themselves with honing techniques to better match new output to existing categorizations or scalings (e.g., DW-Nominate), without much attention to the specific causal or predictive goals of some substantive question.

In this case, the least useful approach would have been to take the ads themselves, run them through some unsupervised algorithm, and then do some descriptive analysis of whatever output resulted. Since we are interested in persuasive effects, one-at-a-time regression was developed specifically to first provide a score for each document, bringing us a step closer to a bottom-up, rather than post-hoc, textual analysis. At this point, the goal is to use that existing document scaling in algorithms that will reveal the words and themes responsible for pro-Kerry and pro-Bush persuasiveness. But how can we be sure that this approach works better than any other subjective, descriptive effort? This is an especially vexed question given that textual analysis generally begins with a bag-of-words approach, where each document is reduced to a vector of word counts (or word proportions), here denoted $\Psi_{k,i}$, where there are $K$ unique ads and $I$ unique words (limited, here, to the top 1000 words, less “stop words” like “the,” “of,” etc.). Why should we think the results of such an simplistic approach should produce anything of empirical utility?

I will present three responses to this question. The first is that simple word frequencies can
do a very good job categorizing even high-level qualities of the sort we are interested in. Table 2 shows the output of a simple supervised categorization of the ads based on a few of the CMAG categories: whether an ad is an attack or a self-promotion; whether it is personal or policy oriented; and whether is is Democrat- or Republican-sponsored. In each case, the supervised procedure is able to categorize over 80% of the ads correctly. Furthermore, the top words from the categories are quite interpretable: dollars, tax, iraq for the attack ads, versus country, believe, healthcare for promote ads; war, country, vietnam for the personal, versus tax, plan, health for the policy ads; health, plan, companies for Democratic ads, versus war, congress, terrorists for the Republican ads. Although there are good reasons to doubt the utility of the CMAG categories, the fact that these procedures can be both predictively useful and quite interpretable should give us confidence in the automated approach.

[Table 2 about here.]

More to the point, are these procedures able to match not the binary scoring of a CMAG category, but the specific scoring we have produced measuring the persuasiveness of each ad? That is, can we find some systematic relation between the words in these ads and their persuasive effect? Again, it is not enough to simply analyze the text of these ads: we must show that such an approach can predict the calculated effects using word frequencies as a proxy for content. The test is to imagine that we do not know the persuasive effect of an ad and then to see how well we can guess its effect, based purely on its textual contents and the contents of the other ads.

There are many possible procedures for such a supervised scaling, where the goal is to generate

---

23 The procedure is detailed in Appendix B. These particular categories were chosen for their importance in studies such as Jacobs & Shapiro (2000), Popkin (1991), and Brader (2006).

24 See Appendix B for details.
\( \hat{\beta}_j \) coefficients that best match the real \( \beta_j \) coefficients associated with each ad. A common approach is to present one method as theoretically preferable, and simply go with that. But as we will see, some procedures work considerably better than others, and some may appear to work best only to fail at harder tests, where apparently weaker procedures excel. Instead, as in developing the one-at-a-time method, I explore a number of approaches, most new to political science, and some devised specifically for the purposes here.

The three classes of supervised scaling utilized here are the following:

**One-at-a-time regression**

\[
\beta = \gamma_i \Psi_{j,i} \quad \forall \text{word } i: 1 \ldots m \\
\hat{\beta}_j = 1/m \sum_{i=1}^{m} \hat{\gamma}_i \Psi_{j,i} 
\]

That is, for each word \( i \), we generate a regression coefficient \( \gamma_i \) by regressing the vector of coefficients from the original one-a-time-procedure on the vector of word proportions for that word in each of the ads. We then generate the predicted values \( \hat{\beta}_j \) by averaging the individual predictions, as before. The one-at-a-time method was devised exactly to suit these situations, where variables are numerous and noisy, regardless of whether those variables are ad broadcast counts, or word proportions.
Spatial (variants: cutoff, weighting, KNN)

\[ w_{j,k} = \frac{1}{1 + \exp(a(\|\Psi_j - \Psi_k\| + b))} \]  

\[ \hat{\beta}_j = \frac{\sum_{k=1}^n w_{j,k} \beta_k}{\sum_{k=1}^n w_{j,k}} \]  

In this case, we take each ad as a position in word space (a simplex in 1000-dimensional Euclidean space, in this case). Each predicted coefficient \( \hat{\beta}_j \) is simply the weighted average of all those ads \( \beta_{-j} \) around it, where the weight is determined by the distances of those ads from the “unknown” one. This weighting function could be logistic with parameters \( a \) and \( b \); or it could be a cutoff, with only those ads within a certain distance being included; or it could be a k-nearest-neighbor algorithm, where only the k closest ads are included. All three variants will be tested here.

Bayesian

\[ P(\Psi_j|L_k) = \prod_{i=1}^n \Psi_{k,i}^{j,i} \]  

\[ \hat{\beta}_j = \frac{1}{n} \frac{\sum_{k=1}^n P(\Psi_j|L_k) \beta_k}{\sum_{k=1}^n P(\Psi_j|L_k)} \]  

Like the spatial approach, this essentially a weighted mean, but in this case, \( \hat{\beta}_j \) is the expectation value given the probabilities that unknown ad \( j \) belongs to each class \( L_k \), where each class \( L_k \) is defined by that ad \( k \). This probability \( P(\Psi_j|L_k) \) is calculated as in equation (9) in the usual “naive” Bayesian fashion, where the probability of a word \( i \) occurring in the class defined by ad \( k \) is simply \( \Psi_{k,i} \), the proportion of that word in that ad, and all word probabilities are taken as independent of each other.
Returning to our second test of the empirical utility of an automated analysis of the ad texts, the simple question is how well these various procedures recapture the original, one-at-a-time-derived coefficients using a cross-validation approach, where each ad’s effect is taken as unknown and estimated based on the texts and coefficients of the other ads. Table 3 shows the correlations between the coefficients estimated using each of these three methods and the original coefficients, as well as the correlations between the various methods. The apparent result is that the word regression works very well, correlating highly with the original coefficients; the Bayesian approach works less well, but still significantly; while the spatial approaches have no significant correlation. Their poor performance is due to the fact that usually only a very few ads are employed in estimating the unknown ad – the maximized cut-off is close, K of KNN is 2, and even the weighting tends to weight distant ads very little. But as the inter-correlations show, these results are not spurious: they correlate with the word regression and Bayesian output significantly, just not with the original coefficients.

[Table 3 about here.]

But although matching the original scaling (or categories) is the usual measure of a supervised approach, examining the correlation between text-estimated and directly-estimated effects is not a true measure of whether we can estimate an ad’s effectiveness from its text. After all, what we are fundamentally concerned with is measuring, predicting, and explaining vote intention; the estimated effect coefficients are only a means to this end. If we were to claim that we could predict the effectiveness of an ad based only on this comparison with original coefficients, a skeptic might

25 Unlike the Bayesian and one-at-a-time approaches, the spatial methods all have one or two parameters to fit, which were selected by maximizing out-of-sample predictions in the third test, below.

26 Although using a Spearman rank correlation test does find significant, if small, correlation between the original coefficients and the hard cutoff and KNN results.
reasonably wonder whether the correlation with direct measurement was close enough to be of any practical use. As before, the real test is whether we can use text to predict ad effects sufficiently well that, when these text-estimated ads effects are employed in out-of-sample prediction, they improve at all on the controls-only baseline. If so, this confirms both the utility of the approach as a tool for campaigns – who might use it to test scripts of potential ads in mid-campaign – and the utility of the text approach in elucidating not just categories such as negativity, but also the subtle and context-specific characteristics that determine an ad’s persuasive success.

The results of this third test of the text approach are presented in the remainder of Figure 2. As before, each line shows the proportion of out-of-sample runs where the ads model out-performs the controls-only baseline. But in these cases, the coefficients are derived from the specified technique, using only the textual similarity between the out-of-sample ad and the in-sample ads. The surprising result is that the one-at-a-time and Bayesian techniques – which best matched the original coefficients – do least well at predicting vote intention, while the spatial approaches all do significantly better than chance. One possible reason for this may be due to differing distributions of $\hat{\beta}_j$ values for each method: while the regression, Bayes, and (to a lesser degree) weighting methods all predict relatively high (near the mean) coefficients for most ads, the KNN and hard-cutoff methods make many fewer “big bets,” assigning most ads a coefficient near 0, and only a few with larger coefficients. The result appears to be that, though the overall correlation with the original coefficients is low, the fewer high bets pay off with better predictive value. But the final line shows that, while the spatial approaches work best, best of all, yet again, is an ensemble of all of the disparate methods: the coefficients generated by averaging all of the text-based predictions work a bit better than even the best of the spatial methods.
Thus far the “prediction” of vote intention has been out-of-sample, but treating the data as pooled. Although there are many fewer data points, we can also predict actually forward in time, taking the in-sample to be the weeks prior to week $t$, and the out-of-sample to be the subsequent week. This is very like what a campaign might use these tool for: testing out unbroadcast (and possibly unproduced) ads to see which ones should have the biggest impact. And again, the proof is in the predicting: Figure 5 shows the success of the predicting vote intention in week $t$ based on a model fit on all weeks $< t$, for each of the weeks leading up to the election. While the coefficients directly estimated using the one-at-a-time method fail to improve on the controls alone, for most of the weeks the text-based coefficients do reliably better than the baseline. (The failure of the directly-measured coefficients is presumably because, in any given week, there are many new ads that have never been run before.) While there are not enough data points here for a proper statistical test, the figure is quite suggestive that these methods can indeed predict the effects of new, untested ads.

[Figure 5 about here.]

But this ability to predict ad effects is more than just a campaign tool: having established that the text of ads can predict both their category types, and their persuasive effects, we are now in a much more secure position to use the textual analysis to infer what substantive characteristics determine those effects. Whatever method we choose, the outcome must be only a few words, since it is very difficult to draw inferences from a ordered list of hundreds of words. But since we will be making inferences based on words with extreme values, it is even more important than with the ad coefficients that those words not be extreme due to chance. Especially with so many words, false negatives are much preferable to false positives, so the job will be to reduce spurious words are
Since there is no direct way to test word scalings out-of-sample in this case, we must rely on the
text prediction results for guidance. As we saw there, the different approaches worked to varying
degrees, but the ensemble of all the methods worked best of all. So that is the approach taken for
the word scaling. Each of the three classes of methods tested before – one-at-a-time, spatial, and
bayesian – has an analogous method for scaling words, most of which are straightforward extensions
of the text-based coefficient prediction.

One-at-a-time

This simply uses the coefficients assigned to words from equation 1. Conveniently, the regres-
sion procedure also provides standard errors, the estimation of which is essential if we wish to avoid
large-error words.

Spatial

For this approach, I adapt the unsupervised scaling developed by Monroe & Maeda (2004) and
Slapin & Proksch (2008), turning it into a supervised method. In the original formulation, inspired
by item response theory, words and documents are presumed to have positions in a one-dimensional
space where (at least in the original formulation) we assume words are drawn from documents in
proportion to the distance between the word and the document. The modification here is that the
documents are presumed to already have positions – their one-at-a-time coefficients – and only the
words \( i \) will be positioned \( (x_i) \) to minimize the following likelihood:

\[
\arg \min_{x_i} \Pi_{i,k} \left( e^{-(x_i-\beta_k)^2} - \Psi_{k,i} \right)^2
\]  

(10)
That is, we minimize the squared errors between, on the one hand, the distance between $x_i$ and the ad position $\beta_k$, and on the other, the proportion of word $i$ in document $k$, $\Psi_{k,i}$. I then employ bootstrapping (on the ad/observation level) to estimate errors.

Bayesian

The approach here is similar to that in equation 9: we want the ad effect expectation given word $i$. Making the usual naive independence assumptions, this is approximately

$$E[\beta|w_i] = \frac{\sum_k \beta_k \Psi_{k,i}}{\sum_k \Psi_{k,i}}$$

(11)

That is, the word score is the expected coefficient given that word, or the sum over ads of each ad coefficient times the probability of that ad given that word. Once again, errors can be estimated via bootstrapping.

But how do we combine the results of these three scalings? They all correlate at about 0.4 – close but not exactly. Simply averaging the three scalings is one approach (after standardizing all three), but this is still vulnerable to the effects of spurious outliers, and makes no use of the error information. To make use of that information, each word score could be calculated via Bayesian updating: knowing the mean and variance of three observations and assuming normalcy, we can easily calculate the updated mean and variance. But this runs into a different problem: the Bayesian word scores have much smaller (normalized) standard errors than the other two procedures, causing the Bayes score to dominate the other two. So instead a much more direct, if less Bayesian, approach is taken for combining these three scores and presenting the top words: for

---

27 This is akin to Bayesian Model averaging, a more general ensemble method for combining disparate models. See (Madigan, Raftery, Volinsky & Hoeting 1996)

28 The calculation resembles the Bayesian shrinkage calculation in equation 3
each list, only the words at least two standard errors from zero are kept, and then of this subset, only words that appear in all three lists are preserved. This is a strong guard against spurious words and inferences.

Table 4 shows the words that appear in all three of the top 30 most effective pro-Democrat or pro-Republican lists, roughly in order of score size. Now we can finally address the substantive questions raised at the beginning of the paper. First, it is notable that there are many more strong pro-Democrat words than pro-Republican. This is because the Democratic ads are in general more effective than the Republican ones; the highest-scoring Republican words are more likely to be there by error, and thus there is less overlap in the top scoring words from the two methods. We can also see that the two word sets imply differing strategies: the Democratic words primarily concern healthcare, a framing strategy where they emphasize their strengths in this area (and possibly the existing policy weaknesses), as well as information about prescription drugs. The Republican words indicate, on the other hand, a more negative, attack strategy – not a different set of agenda items, but an entirely different mode of persuasion. The top word is “Democrat,” presumably meant pejoratively as both a noun and an adjective. “Deficits” may or may not be an attack word, depending on your view, but the “boat veterans” part clearly picks up on the famous Swift Boat Veterans for Truth ads – confirming that these results are consistent with existing views about the 2004 campaign. Finally of note is the perhaps surprising presence of “Michael Badnarik,” the libertarian candidate. Traditionally one might think the libertarian would be drawing more Republican votes, but in 2004 there were many previous Bush voters who were now dissatisfied with Bush, and every dissatisfied vote for the libertarian was one fewer vote for Kerry. For both candidates, the presence of the swift boat ads and “moveon” indicates the power of external advertisers, but although MoveOn has a reputation for harder hitting ads (relatively
speaking), these seem not to be the ones that were most effective. Whereas on the Republican side, even the libertarian ads were in a sense negative, inasmuch all of these ads are less directed towards persuading voters to vote for Bush and his policies than they are engaged in persuading voters not to vote for Kerry.

[Table 4 about here.]

Of course, none of this interpretation would be possible without pre-existing expertise on the campaign and its themes. The utility of the bottom-up textual approach lies, rather, in distinguishing which of the myriad strategies deployed actually worked, not just in one ad but in a broad range of ads as a systematic result of some consistent persuasive effect. Unlike merely qualitative opinions, the bottom-up approach establishes statistically which ads are the most effective, and then establishes that the textual contents of those ads – even viewed as simplified word counts – can predict abstract categories and even the effects of those ads. Only having done both of those steps can we be certain that the words deriving from the textual procedures are more than just an automated version of descriptive work. The words reveal which aspects of the ads do the actual persuasion, and show that choosing between affective, rational, and agenda-setting theories is an unnecessary and misguided first step, when the real campaign is fought with an asymmetric mixture of framing, information, affect, and branding.

Each step of the process here has been necessary to get us to these substantive conclusions. We saw that other methods for estimating ad effects using this data failed, whether using human categories or algorithmic variable reduction like PCA or LASSO. Instead, a new technique, one-at-a-time regression, was shown through out-of-sample validation to measure and predict ad effects far
better than the other methods, and also to provide estimated effects for every ad. This latter quality was necessary for the second step, explaining why the ads vary in their effects, and inferring the successful persuasive strategies underlying those effects. But before plunging into text analysis, an important methodological step here is showing that the automated text procedures developed here actually work, in the same way the one-at-a-time regression was shown to work: by predicting vote intention. New techniques were developed that allow us to predict ad effects based purely on their textual similarity to directly measured ads, both validating the textual approach and establishing a potentially powerful campaign tool. Finally, having established the empirical validity and utility of these text procedures, related word-scaling methods were developed and aggregated to give us the best, most confident set of words characterizing the most effective pro-Democratic and pro-Republican ads. From these words it is immediately obvious that the most effective strategies are very different between the two sides, an insight that would be difficult with more traditional methods. These steps – one-at-a-time regression, out-of-sample testing, out-of-sample text-based prediction, and finally word scaling – together constitute a more general bottom-up procedure for demonstrating and understanding the persuasive effects of linguistic data in a wide variety of contexts where actors attempt to persuade each other with speech.
Appendix A: Omitted variable bias?

One might reasonably suspect that the one-at-a-time approach would be hindered by omitted variable bias – after all, each regression is omitting hundreds of variables!

Specifically, if we know that the true model is:

\[
Y = X\beta + Z\delta + \epsilon \tag{12}
\]

where \(X\) and \(Z\) are two sets of independent variables that together constitute \(Y\), and we instead regress \(Y\) on \(X\) alone, we get \(\hat{\beta}\):

\[
\hat{\beta} = (X'X)^{-1}X'Y
\]

\[
= (X'X)^{-1}X'(X\beta + Z\delta + \epsilon)
\]

\[
= \beta + (X'X)^{-1}X'(Z\delta) + (X'X)^{-1}X'\epsilon
\]

\[
= \beta + (X'X)^{-1}X'(Z\delta) \tag{13}
\]

assuming iid errors. Thus our estimate for \(\hat{\beta}\) has a bias due to regressing \(Y\) on only the \(X\) variables, equal to the last term in (8). In general, one can say little about the magnitude or direction of this error. But say we believe that all the ad variables in the set of 359 should be included, but regress \(Y\) on only one of them at a time. Then each \(\hat{\beta}\) will be biased by the omission of all the other ad variables, but in a predictable way. Specifically, for the \(j = 359\) individual ad regressions we get
for the ad betas (omitting the controls):

\[ \hat{\beta}_1 = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'X_3\beta_3 + \ldots \]

\[ \hat{\beta}_2 = (X_2'X_2)^{-1}X_2'X_1\beta_1 + \beta_2 + (X_2'X_2)^{-1}X_2'X_3\beta_3 + \ldots \]

\[ \hat{\beta}_3 = \ldots \]

\[ \vdots \]

\[ (14) \]

or

\[ \hat{\beta}_i = \sum_{k=1}^{j} \hat{\psi}_{i,k}\beta_k \]

\[ (15) \]

where \( \hat{\psi}_{i,k} \) is the coefficient from regressing \( X_i \) on \( X_k \). So if we assume that the \( j \) variables are everything (a strong assumption, of course), then we can solve for the true \( \beta_k \) coefficients: since we can calculate all the \( \hat{\beta}_i \) and \( \hat{\psi}_{i,k} \) coefficients, solving for the \( \beta_k \)s is just a matter of \( j \) equations and \( j \) unknowns.

Alas, practicalities get in the way at this point. With 359 simultaneous linear equations and a matrix of coefficients with a determinant on the order of \( 10^{-500} \), this system cannot be solved analytically. Ironically, the best solution is found (say, by Mathematica or Matlab) by simply carrying out the multivariate regression, regressing the vector of \( \hat{\beta}_i \) on \( j \) \( \hat{\psi}_{i,k} \) vectors. Using that method, we can find a fairly close (if approximate) solution to the system, resulting in a set of “true” \( \beta_k \) values.

Returning to the original question now, does this “correction” for omitted variable bias help? The answer is no. When we test this procedure out of sample, only 63% of the runs do better than the controls alone. This shows that the procedure is not junk, since as we have seen, most
meddling with the beta values will produce vastly lower OOS percentages. In fact, the “corrected” betas are usually quite different from the original one-at-a-time beta-hats, yet they still work fairly well. It’s just that regressing each ad one at a time and averaging the final prediction works better, regardless of whatever omitted variable bias that may introduce. In fact, it may be that “correcting” the betas merely reintroduces some of the same problems of over-fitting that the one-at-a-time method sought to avoid – as the very low full-sample RMSE suggests. The power of the one-at-a-time-then-averaging method derives precisely from omitting all the other variables: when there are so many variables, the bias due to omission is greatly exceeded by the gains made in avoiding over-fitting.
Appendix B: Predicting ad categories

The algorithm employed categorizes an ad by whichever set of ads – category A or B – has the lowest root mean square distance:

\[
\text{Cat}\{X\} = \min \left\{ \frac{1}{|A|} \sum_{j \in A} \sqrt{\frac{1000}{\sum_{i=1}^{1000} (w_{x,i} - w_{j,i})^2}}, \frac{1}{|B|} \sum_{j \in B} \sqrt{\frac{1000}{\sum_{i=1}^{1000} (w_{x,i} - w_{j,i})^2}} \right\}
\]

(16)

That is, the unknown ad X gets category A or B by simply adding up and averaging the squared differences between the word proportions in ad X, and those in each of the members of A and B.

To create the words listed in Table 2, rather than simply take the most frequent words (largest dimensions) from that vector, the vector of all words (ie, created from both categories) is subtracted from the category vector, to yield those words that are most distinctive of that category. The results, as Table 2 shows, are not simply the words we might assume would be obvious flags of attack-ness or personal-ness or Democratic-ness. Rather, they are substantive words, whose ability to define an ad as an attack or a policy ad comes not from any simple one-to-one mapping between a word and a category, but rather from the aggregate distributions of words which collectively constitute an attack or a promoting ad. But nor are the words inexplicable, or even very surprising: attacks ads feature jobs, taxes, Iraq, health, dollars, oil, and so on – an interesting mix of Republican and Democratic concerns, but collectively identifying nevertheless. Promotion ads feature more uplifting and general concepts, such as believe, world, family, life, strong. Personal vs Policy might seem an even more abstract dichotomy, though one that is fundamental to questions of whether ads merely tug the emotions, or in fact make substantive arguments. But here the algorithm does even better, even though the word sets themselves have more overlap. Again, these categories are
particularly challenging because of the differing concerns of the two parties (e.g., Vietnam and boat, which pertain exclusively to Kerry) that are all mixed together. But one can note the more diffuse content of the personal, and the over-riding issue of the war; while the policy category appears more domestic and specific: health, plant, companies, million, insurance, drug, dollars, etc.

Finally, with the party identification challenge (the truth of which, unlike the other categories, is quite objective), the algorithm again does quite well, with more domestic and budgetary concerns on the Democratic side, and taxes, terrorists, freedom and liberals on the other. The comparisons between these dichotomies and the other two are also interesting: the Republican words tend to be more oriented towards attacking threats – either from terrorists or liberals – while the Democratic words tend to be more (domestic) policy oriented; on the other hand, the attack words themselves seem to reflect Democratic concerns. It would seem that the out-of-power party engaged in more attacks, but the in-power party was equally dedicated to threats – though perhaps their attacks were more ambiguous. Or it may simply be that, as we saw earlier, the Democrats simply ran many more ads, so that the majority of attack ads might reflect their concerns, while there were also an abundance of domestic policy ads as well.
References


Figure 1: Out-of-sample RMSE (minus noise) with simulated data; lower is better; x axis denotes increasing noise in generating DV.
Figure 2: For each out-of-sample test, a model is fit on a random 95% of the observations, and then the dependent variable, vote intention, is predicted out-of-sample and compared with the actual measured vote intention. The baseline model predicts vote intention using only demographic variables (such as age, sex, party ID, etc). Each model adds on top of the demographic variables a method for measuring advertising effects. The more runs a model beats the baseline, the better it is able to utilize the ad data, and the better it measures their effects. Longer bars are better; bars outside the red-green band < 0.05 probability by chance.
Figure 3: Distribution of effect coefficients for Democratic- and Republican-sponsored ads.
Figure 4: The predicted effect on vote share, if one party buys only the most effective (highest coefficient) ads. Errors are bootstrapped; shrunk coefficients used.
Figure 5: For each week, the model is fit on all the preceding weeks, and then used to predict vote intention for that week. “Real coefs” uses the one-at-a-time estimated ad coefficients, where as the other two models predict vote intention for that week using coefficients estimated using the textual similarity between those ads and the ads whose effects were measured in the preceding weeks. The text-based approaches do better than nothing, and better than using the real coefficients, since in a given week, many ads are new and have never run before. Figure shows RMSE performance relative to the controls-only prediction; higher is better.
Table 1: Assessing the relative effects of the independent variables (including the all-ad variable) on Democratic intended vote share.\(^a\)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta at 30th Pctile</th>
<th>Beta at 70th Pctile</th>
</tr>
</thead>
<tbody>
<tr>
<td>party ID</td>
<td>0.128</td>
<td>0.120</td>
</tr>
<tr>
<td>ideology</td>
<td>0.020</td>
<td>0.019</td>
</tr>
<tr>
<td>age</td>
<td>0.014</td>
<td>0.013</td>
</tr>
<tr>
<td>unemployed</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>black</td>
<td>0.012</td>
<td>0.013</td>
</tr>
<tr>
<td>religious</td>
<td>-0.019</td>
<td>-0.019</td>
</tr>
<tr>
<td>south</td>
<td>-0.011</td>
<td>-0.010</td>
</tr>
<tr>
<td>urban</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td>all-ad var(^a)</td>
<td>0.040</td>
<td>0.020</td>
</tr>
</tbody>
</table>

\(^a\) The effect size reported in each cell is simply the change in the dependent variable (Democratic intended vote share) produced by the specified change in independent variable.

\(^b\) The unified ad variable is \(\hat{y} = \frac{1}{j} \sum_j \hat{\beta}_j X_j\), based on equation (6). That is, it is the average of the product of each ad count variable times its estimated coefficient (ie, the average prediction). It is substantively meaningless, but gives a single variable to assess the overall impact of the ads.

\(^c\) The “clarify” package in Stata was used to estimate the effect on the dependent variable when an independent one is shifted from its 25th to 75th percentile, all other variables being kept at their means.
Table 2: Top words for each CMAG category along with the success rate of the text-based classification technique.\(^a\)

<table>
<thead>
<tr>
<th>Attack Promote</th>
<th>Personal Policy</th>
<th>Democrat Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>jobs country</td>
<td>war jobs</td>
<td>jobs voted</td>
</tr>
<tr>
<td>dollars believe</td>
<td>country tax</td>
<td>plan taxes</td>
</tr>
<tr>
<td>tax jobs</td>
<td>vietnam plan</td>
<td>tax war</td>
</tr>
<tr>
<td>iraq people</td>
<td>believe health</td>
<td>health times</td>
</tr>
<tr>
<td>voted healthcare</td>
<td>served people</td>
<td>country congress</td>
</tr>
<tr>
<td>national plan</td>
<td>people companies</td>
<td>american people</td>
</tr>
<tr>
<td>taxes american</td>
<td>truth million</td>
<td>time tax</td>
</tr>
<tr>
<td>war health</td>
<td>life iraq</td>
<td>companies terrorists</td>
</tr>
<tr>
<td>million world</td>
<td>time insurance</td>
<td>healthcare believe</td>
</tr>
<tr>
<td>care family</td>
<td>world care</td>
<td>iraq health</td>
</tr>
</tbody>
</table>

\(^a\)The percentage is of those ads that belong to one of the two paired categories that were correctly assigned to the true CMAG category. For all three pairings, most ads had been put in one of these two categories; those that were not, were not included in the testing. Words are listed in descending absolute value of score.
Table 3: Correlations between the ad coefficients assigned by the various text-based estimation methods.$^a$

<table>
<thead>
<tr>
<th></th>
<th>real</th>
<th>word reg</th>
<th>bayes</th>
<th>cut-off</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>word regression</td>
<td>0.841*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bayes</td>
<td>0.266*</td>
<td>0.520*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hard cut-off</td>
<td>-0.004</td>
<td>0.225*</td>
<td>0.091</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>0.014</td>
<td>0.203*</td>
<td>0.204*</td>
<td>0.538*</td>
<td></td>
</tr>
<tr>
<td>weighting</td>
<td>-0.006</td>
<td>0.277*</td>
<td>0.198*</td>
<td>0.940*</td>
<td>0.681*</td>
</tr>
</tbody>
</table>

$^a$ Values with * significant at $p < 0.05$. 
Table 4: Most effective words for shifting intended vote in pro-Democrat and pro-Republican directions.

<table>
<thead>
<tr>
<th>pro-Democrat</th>
<th>pro-Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>hospital standards</td>
<td>democrat deficits</td>
</tr>
<tr>
<td>jail blocking mess</td>
<td>michael badnarik</td>
</tr>
<tr>
<td>canada broken strength</td>
<td>prices</td>
</tr>
<tr>
<td>moveon</td>
<td>boat veterans</td>
</tr>
<tr>
<td>cost medecare</td>
<td>profits</td>
</tr>
<tr>
<td>negotiating</td>
<td></td>
</tr>
</tbody>
</table>