Campaign News and Vote Intentions

Analyses Using Manual and Automated Coding
of Sentiment in Canadian Newspaper Content

Stuart Soroka Lori Young Marc André Bodet

McGill University

Paper prepared for presentation at the Annual Meeting of the Canadian Political Science Association, Vancouver BC, June 4-6 2008. We are grateful to Mark Pickup for providing the dataset of Canadian election polling results, to Blake Andrew for his help developing and managing the manual content analyses, to Mark Daku for his ongoing work programming *Lexicoder*, and to John Galbraith for comments on a previous draft. This work was funded in part by the Fonds québécois de la recherche sur la société et la culture, the Donner Canadian Foundation, the McGill-Max Bell Strategic Initiative, and the McGill Institute for the Study of Canada.

That mass media are central to modern representative democracy requires little discussion. Mass media play an important role in producing an informed (or at least moderately informed) public. They are critical to the dissemination of information about, for instance, national conditions, government activities, and public policy issues. Media content is accordingly strongly connected with public opinion, politics and policy. Indeed, there are burgeoning literatures detailing the relationship between media content and, for instance, public attentiveness to issues, policymakers' framing of policy matters, and public attitudes about public policy.[1]

Whether the role of mass media is to lead or to follow is of course in many cases not clear; it is likely that at any given time mass media are doing a bit of both. Media content can be regarded in two often empirically inseparable ways: (1) it can reflect the issues, themes and actors that are currently prominent in public debate, and (2) it can be a potential driver of public opinion and policy. In the former case, mass media act simply as mirror. Media content in this view is a useful summary indication of the more general public sphere. In the latter case, mass media are not mirroring but affecting. Media content differs from what citizens or politicians currently think, and has the potential to affect these actors' attitudes.

Both of these connections between media content and public opinion may be particularly strong during election campaigns. These are periods of heightened attentiveness to political issues for journalists, for many citizens, and for policymakers. The modern election campaign is hugely dependent on media – on advertising in many cases, but on news content as well. Indeed, election campaigns are almost by definition *media* campaigns: for most citizens, media are the principle if not the only source of information about leaders, candidates, parties, policies, and, of course, the horserace.[2] Journalists are highly attentive to the campaign as well as to the state of public opinion, particularly vote intentions. And many citizens are likely, compared to other periods at least, more attentive to media content. Given this heightened mutual attentiveness, we might expect the link between media and the public during an election campaign to be especially strong.

This paper explores this relationship between campaign-period media content and public opinion, focusing on two recent Canadian federal election campaigns. Specifically, we examine the relationship between the tone of media coverage and vote intentions for the major parties over the 2004 and 2006 Canadian federal

---

[1] On attentiveness see, e.g., McCombs and Shaw 1972; Behr and Iyengar 1985; Soroka 2002; on framing see, e.g., Iyengar 1996; Baumgartner and Jones 1992; on policy attitudes see, e.g., Hall-Jamieson and Cappella 1998; Fan and Norem 1992; Soroka 2003.

[2] Horserace coverage has received a particularly large amount of attention in the literature. See, e.g., Craig 2000; Fletcher 1991; Graber 1976; Jaimeson 1992; Mendelshon 1993; Patterson 1993; Wilson 1980.

election campaigns. Results suggest that this relationship is rather strong – that, due to a number of possible mechanisms, media content and pubic opinion are positively related. We also explore a more specific (and perhaps more ambitious) possibility, however: With a good measure of tone in media content, is it possible to actually *predict* vote intentions?

It looks like maybe we can. That this is true suggests, at a minimum, the potential importance of media content in understanding election campaigns. It may also demonstrate a rather striking effect of media content on vote intentions during election campaigns. And note that while demonstrating a link between media content and public opinion during campaigns is by no means an original contribution,[3] that media content analytic data can actually be used as a (powerful) leading indicator for trends in opinion is.

So too is the use of automated content analytic procedures. Automated content analysis is a powerful tool to compute relevant indicators systematically, for large quantities of data. To date, most content analyses in political science have been performed manually. Reliability and homogeneity in the coding process are often problematic, particularly for comparatively subjective codes such as "tone." Automated content analysis goes some way towards solving this problem. It also makes feasible the (identical) treatment of a previously unimaginable quantity of data. In addition to exploring the strength of the link between media content and vote intentions, then, this work seeks to make a methodological contribution – namely, we apply and test both (finished) manual and (preliminary) automated systems of coding to large bodies of campaign-period media data. Previous work in political communications has typically worked with much smaller samples; and while it has automated the coding of topics the automation of tone has met with less success. Our comparatively simple method seems to work reasonably well, and can be easily applied to other bodies of data, using a forthcoming, freely available, Java-based (multi-platform) software called *Lexicoder*.[4]

Data and results are presented in two sections below. The first describes the manually-coded data for 2004 and 2006, and explores the strength of these data in predicting vote intentions in these two campaigns. The second follows the same structure, but uses automated content analytic data for 2006. Overall, results speak

---

[3] There are of course vast literatures on the link between media and opinion in campaigns. For work in Canada see, e.g., Blais and Boyer 1996; Mendelsohn 1994, 1996; Mendelsohn and Nadeau 1997; Johnston et al. 1992; Wagenberg et al. 1998. For work elsewhere see, e.g., Brians and Wattenberg 1996; Druckman 2004; Krosnick and Kinder 1990. For a more thorough review of the earlier US literature, see Weaver 1996.

[4] *Lexicoder* is the product of ongoing work by Young and Soroka. The first version is being programmed by Mark Daku, using the source code from *Yoshikoder* by Will Lowe. A public version of *Lexicoder* – the functioning software, and the source code, and a selection of topic- and tone-oriented dictionaries – will be available in Fall 2008 from the Media Observatory website, http://www.media-observatory.mcgill.ca.

to the potential for automated coding in the study of political communication, and to the relationship between media content and opinion in election campaigns.

Manually-Coded Media Data and Vote Intentions

Data

Our first analyses rely on a two bodies of data. The first is relatively simple: a database of all commercial polls in each of the five Canadian federal election campaigns from 1993 to 2006. There are over 300 polls in total, but a good many more polls in the more recent elections. In 2006, there were 8 pollsters regularly in the field, with over 40 regular polls and two rolling-cross sections of a few hundred respondents per day conducted by SES and The Strategic Council. In 1993, in contrast, there were 6 pollsters in the field but just 13 polls over the 41-day period. Indeed, polling before 2004 was sporadic enough that we cannot attempt a parametric analysis in those years. We accordingly focus here on just the two elections in which polls are most prevalent – the 2004 and 2006 elections.

The other database is a body of content analytic data, manually coded by expert coders during the 2004 and 2006 federal campaigns.[5] The 2004 and 2006 studies were conducted separately, but are directly comparable methodologically speaking. Each tracks all campaign content – news, editorials, and opinion pieces – published in seven major daily newspapers across Canada (five English-language, two French-language): *Vancouver Sun, Calgary Herald, Toronto Star, Le Devoir, La Presse, National Post and The Globe and Mail*. Daily coding for the 2004 election began one week before the writ dropped, lasting for six weeks until the day of the election, June 28, 2004. The 2006 campaign – among the longest in Canadian history – accounts for eight weeks of coding material before the election on January 23, 2006. In total, 6694 articles are included in this dataset: 4,280 news stories and 2,414 editorial and opinion pieces.

In both 2004 and 2006, coders surveyed the main news sections of the major Canadian dailies for the duration of the campaign. There were about a dozen coders for each campaign, introduced to the study during formal training sessions that included a series of practice coding exercises and a guide for our online data entry system. Coding happened daily, as the campaign progressed. Coders were responsible for a different newspaper each week, in order to test for (and avoid) any coder effects or bias. Stories were also randomly selected for double-coding throughout the campaign to check inter-coder reliability – the consistency with which different coders come up with identical codes. (All measures included in this analysis achieved an appropriate level of reliability. Detailed methodological information is available at the Media Observatory website.)

---

[5] These data are distributed by the Media Observatory of the McGill Institute for the Study of Canada (http://www.media-observatory.mcgill.ca). The description here borrows considerably from Soroka and Andrew, N.d.

Coding captured a body of relatively objective data, including mentions of issues, parties, and leaders. Most critical for the forthcoming analyses, coding also included one set of subjective codes for tone – positive, negative or neutral – for parties and leaders. The specific instructions for coding the tone of media content were as follows: the default for all mentions is neutral; a leader or party mention has to be clearly 'good press' or 'bad press' to be coded as such. (Note that is similar to, e.g., Brady and Johnston 1987.) Put another way, unless the story was obviously and intentionally positive or negative, a mention of a leader or party is neutral. This is what you might call latent rather than manifest measurement of election news content — it captures tone evident in the reporting of or commentary on a given event, rather than negativity or positivity of the event itself.[6] For that reason, careful attention was paid to training and to reliability analyses for this indicator.

Note that the tone code does not simply reflect reports of leaders and parties criticizing policy platforms and records of competitors. Our measure of an article's tone, instead, reflects critical and positive commentary of the main leaders and parties from sources other than the main protagonists of the campaign itself. For instance, reporting a Harper speech in which the Conservative leader objected to or attacked something about Paul Martin was considered neutral – just reporting the news. Reporting that speech and using it to further discuss Martin's failings was considered negative, however. Our coders also noted tone when, for example, an economist issues an endorsement for a party's tax policy proposal, but not when another party leader attacked (or endorsed) it. The tone measure also captures assessments of leaders and parties' performance in the campaign and in public opinion polls. Reports of a party "surging ahead" in the polls were noted as a positive. Conversely, stories noting "uninspiring" or "gaffe-prone" campaign performances were duly recorded as negative press. The overall result, then, was that mentions in news stories were predominantly neutral, and mentions in editorial and opinion pieces were mainly negative or positive. To be sure, the coders have disregarded some of the subtle tone conveyed in articles and headlines, but then so do everyday citizens.

Analysis

This combination of polling and media data allow for, we believe, a relatively wide range of analyses on electoral campaigns and political communications. We focus here on one relatively narrow question, however: How well can we predict vote intentions using media content?

We explore this question here using a relatively simple Ordinary Least Squares (OLS) regression model of the current vote share for each of the major two parties, and just two matrices of lagged independent variables. The first is a series of dummy variable for polling firms in our sample, including SES, The Strategic Council, Ekos, Léger Marketing, Decima, Ipsos Reid, Environics, Zogby Poll, Pollara, and Compas. These

[6] See Andrew 2007; Riffe et al. 2005.

variables are equal to 1 if a given firm has a poll on the field on a given day and 0 otherwise; they are intended to capture 'house' or 'pollster' effects – the tendency for different firms to capture slightly but systematically different distributions of vote intention, due to methodological decisions relating to, for instance, the partition of undecided voters and "don't knows". (See, e.g, Jackman 2005; Pickup and Johnston 2007; McDermott and Frankovic 2003.) Note that these pollster effects can be viewed as one of several different types of error in polling data, alongside what classical statistics labels 'random error,' a function of sample size, clerical errors, etc. And while this random error is typically handled by using least squares estimates, the error resulting from pollster effects is dealt with by using the dummy variables outlined here.[7] In any case, we do not interpret pollster effects dummies below, but include them only as controls.

The second matrix of independent variables includes four net tone measures, for (1) the Liberal Party, (2) the Conservative Party, (3) Paul Martin, and (4) Stephen Harper. 'Net tone' is simply the number of positive mentions for a leader or party minus the number of negative mentions on a given day.[8]

Put more formally, our prediction model is as follows,

$$Vote_{p,t} = a + \sum(\beta_f * Pollster_{f,t}) + \sum(\omega_\eta * Tone_{\eta,t-k}) \; [+ \omega_\eta * Vote_{p,t-k}] + \epsilon_t \qquad ,$$

where polled vote intentions for party (p) at some time (t) is a function of a set of dummy variables capturing pollster effects for each firm (f) at that time and a set of net tone measures for each major party and leader (η) lagged by k time periods. We also include in square brackets above the dependent variable, $Vote_p$, also lagged k time periods. On the one hand, while a model's predictive capacity can be can be limited by not taking into account a strong autoregressive (AR) process (the tendency for a value at t to be strongly related to a value at t-k), excluding the AR process provides a good opportunity to evaluate the predictive capacity of media content variables without the strong input of the lagged dependent variable. On the other hand, any effort to produce estimates of some variable sometime in the future would certainly include, if available, recent or current values of that variable. Since we do have current polling results at any given (t-k) point in the campaign, there is nothing preventing us from including the party's vote share in the prediction. Indeed,

---

[7] Note also that there is another aspect of error in polling results that is rarely discussed but that is often implicitly accepted in public opinion research. Because respondents may be in a better position to accurately express vote intention later in the campaign, survey responses may be more reliable closer to the election date. As a consequence, time series of vote intentions may be prone to temporal heteroskedasticity – random error variance may not be constant across time. This violates a critical assumption of OLS, but we do not address it here. We simply assume temporal homoskedasticity.

[8] Note that for 'net tone' we lump together all articles from all newspapers – we do not give newspapers different weights based on audience reach, nor do we distinguish between the potentially different content in different newspapers. While newspapers differ in levels of tone for different parties (Soroka and Andrew, N.d.), however, they follow very similar trends over the campaign. It is not clear that there is much to gain by looking at newspapers separately.

including both media and lagged vote share together provides a strong test of the degree to which media content improves the prediction, above and beyond what we would know using just the current vote share; and a model using *just* lagged vote share provides a good baseline against which to measure the contribution of media variables. We accordingly estimate each model three ways below: (1) vote share at *t-k*, (2) media content at *t-k*, and (3) both vote share and media content at *t-k*.

The choice of lag – that is, the value of k – is driven by a combination of pragmatic and statistical considerations. Pragmatically speaking, the further back the lags are (the greater the value of k), the further forward we are able to predict. A model using media at *t-1* would allow us to predict only one day forward; a model using media at *t-6* would allow us to predict six days ahead. At the same time, with a limited election period high order lags are costly in terms of degrees of freedom – each additional lag means one less data point. We would thus ideally select lags over some kind of middle period, not too proximate (so we can predict), but also not too distant (to preserve sample size).

That said, the data will also speak for themselves. If there is a relationship between media content and vote intentions, the strength and timing of that relationship should be relatively clear in preliminary models and in simple cross-correlations, calculated between media content and vote intentions at various lags and leads. Our preliminary tests (not shown here) suggested that correlations were typically strongest at lags *t-4* through *t-6*. This finding was roughly consistent for both Liberal and Conservative vote intentions, in both the 2004 and 2006 elections, and using any combination of the net tone measures. The lags also have the happy coincidence of allowing us to predict four days in advance, and not losing too many degrees of freedom. We accordingly include lags *t-4* through *t-6* of media content in all our estimations of equation 1. For vote share, where included, we use just *t-4* – the most proximate polling data we would have if we were mid-campaign, using today's media content to project vote share four days ahead.[9]

[Tables 1 and 2 about here]

Results from those estimations are included in Tables 1 and 2. With three lags of four different variables, each of which is correlated to the others, the model is vastly over-specified. Coefficients can be very difficult to interpret, given that there will be several coefficients capturing what is, substantively speaking, a single effect. Standard errors will also be inflated by multicollinearity so coefficients will rarely achieve common levels of statistical significance. This is of course common for prediction models that seek to explain as much variance as possible, and place little emphasis on interpreting individual coefficients. It presents no particular problem for our work here, but it does mean that we should not place too much weight on the individual coefficients.

---

[9] Preliminary tests confirmed that using a single lag for vote intentions was all that was required – once vote intentions at *t-4* are included, vote intentions at *t-5* and *t-6* have no significant effects.

In an effort to make results more readily interpretable, the first row of Tables 1 and 2 shows the summed coefficients (and related standard error) for all six lags of tone relating to the Conservative Party and Stephen Harper. The same is done for the Liberal Party in the second row. Each provides an omnibus test of the direction and magnitude of the effects of tone for the two major parties. The rather obvious expectation is that Conservative tone will be positively related to Conservative vote intentions and negatively related to Liberal vote intentions, while Liberal tone will be positively related to Liberal vote intentions and negatively related to Conservative vote intentions.

Let us begin with result for the Conservatives in 2006, the first three models in Table 1. Column 1 shows results using just lagged vote share and pollster effects. The R-squared is very high, .895 – already almost 90% of the variance in vote share at *t* is explained using just vote share at *t-4*. But note that while the R-squared provides a summary of the proportion of variance in *y* explained by *x*, it does not provide the piece of information we are most interested in where prediction is concerned – exactly how close are the predictions to the future values of *y*? To better assess this critical piece of information, we rely here on the Mean Absolute Error (MAE), which captures the average gap between the prediction and the actual vote intentions.[10] For this first model, the MAE is .952. (Note that the MAE is in the same unit as the dependent variable, so we are talking here about a prediction that on average misses the Conservative vote share by just less than one percentage point.)

Column 2 shows results for the 2006 Conservative vote share, this time using just lagged media variables. First, note that Conservative net tone is positively related to Conservative vote share, and Liberal net tone is negatively related to Conservative vote share, as we would expect. The MAE almost doubles, to 1.85, but we should not lose sight of the fact that this is a model that uses *just* media content. And the MAE is at its lowest in the third column, combining lagged media and vote shares. That said, the improvement over the first model is rather slight, and there is no clear effect of either tone variable once lagged vote shares are included.

Media effects are clearer in the model for Liberal vote share. The MAE is considerably lower for the combined model than for the model using just vote share – .468 versus 1.103. Using media content in the prediction of Liberal vote share cuts the average error in half. And in both columns 5 and 6, the effects of Conservative and Liberal tone are corrected signed.

Table 2 presents roughly similar results for 2004. The accuracy of both the Conservative and Liberal predictions improves considerably with the inclusion of the lagged media variables. Liberal tone seems to make little difference to the Conservative vote share in 2004, but Conservative tone clearly matters. Indeed, it is Conservative tone that matters for the Liberal vote share too – good Conservative coverage seems to be related to decreasing Liberal votes.

---

[10] On the value of the MAE and SEE a goodness of fit measures in prediction and forecasting, see Krueger and Lewis-Beck 2005.

That predictions improve with the inclusion of media content is strong evidence of the effect of campaign-period media, we believe. The degree to which media both reflect current vote shares and affect future trends is further evidenced by the strength of models including just media content. Looking across Tables 1 and 2, it is striking that we can explain about 70% of the variance in major party vote intentions using just lagged media.

[Figures 1 and 2 about here]

The predictive power of media content is further illustrated in Figures 1 and 2, which show both polls and (lowess-smoothed) predictions for the two major parties over the 2004 and 2006 elections, relying on the models that include media content only. In both cases, predictions track results rather well. There are some interesting exceptions, including the beginning of the last week of the 2006 campaign and last few days of the 2004 campaign. We return to these exceptions below. First, however, we re-estimate these same models using a different body of media data.

Automated Media Data and Vote Intentions

Data

The second body of content-analytic data is an entirely new database of all election stories from the last five Canadian federal campaigns. The database includes the five English-language newspapers in the manually-coded sample, plus the *Montreal Gazette*. French newspapers are not included, since the automated analysis can only work with English text. The *Gazette* achieves a little more regional variance, but clearly is no substitute for the two French-language newspapers. We should thus keep in mind that this sample is not regionally representative in the same way as the manually-coded sample.[11]

Mentions of parties and party leaders can be captured relatively easily in an automated analysis – even standard database software can record the number of mentions for different words in a corpus. Our automated analysis begins, then, by identifying all those sentences in which either of the major parties or leaders are mentioned. We then automate the coding of tone, using a relatively simple "bag-of-words" approach.

The software extracts, for instance, all sentences including a reference to Stephen Harper, and then counts the number of positive and negative words in those sentences. This is a relatively simply proximity-based process, then – that is, a process that relies on proximity (or local co-occurrence) of affect words and the "subject" of interest in a text to improve sentiment analysis relative to full-text

[11] We should also note that, even for the five English-language newspapers in both databases, the samples will not be identical. Manually-coded data were gathering during the campaigns, from hard copies of newspapers. Automated data were gathering from full-text indices in Nexis. There is of course a good degree of overlap, but there are bound to be some differences as well.

analyses. (See, e.g., Pang, Lee and Vaithyanathan 2002; Mullen and Collier 2004.) The resulting 'net tone' measure in this case is the number of positive words, minus the number of negative words, as a proportion of all words in sentences mentioning Harper (or Martin, or Conservatives, or Liberals).

Defining positive and negative words is of course no small matter. There are a good number of established content analytic dictionaries focusing on sentiment (or valence, or tone) in text. These dictionaries vary widely with respect to valence categories and scope; there is also surprisingly little overlap among dictionaries and – to the extent that there is – there are many discrepant codes. To improve the scope of coverage and the consistency of valence codes, we merged nine of the most commonly used affective lexical resources.

Three of the nine dictionaries include positive and negative categories. The General Inquirer (Stone et al. 1996) includes two valence categories labeled "Positiv" and "Negativ" (n=4295); WordNet-Affect 1.1. (Strapparava and Valitutti. 2004) labels a subset of affective words from WordNet "positive" or "negative" (n=1640); and the word list used in the TAS/C text analysis software (Mergenthaler 1996; 2008) is labeled for positive or negative "emotion tone" according to the dimensions of pleasure-displeasure, approval-disapproval, attachment-disattachment and surprise (n=4058).

Several dictionaries contain relevant affect or emotion categories, but are not coded for valence. These were manually labeled, omitting ambiguous categories. From the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al. 2001) we included the positive category "Positive emotion" and the negative categories "Negative Emotion," "anxiety," "anger," and "sadness" (n=1502). From the emotion class of the Regression and Imagery Dictionary (RID) (Martindale 1975, 1990) we include the positive categories "positive affect" and "glory" and the negative categories "chaos," "aggression," "diffusion," "anxiety" and "sadness" (n=1056). The Roget's Thesaurus (Roget 1911) was also manually labeled. Its categories are numerous, but include, for instance, positive categories such as "benevolence," "vindication," "respect," "cheerfulness" and "intelligence" and negative categories such as "insolence," "malevolence," "painfulness," "disappointment," and "neglect" (n=47596).

Finally, several dictionaries classify words along scales of valence and were used on a discretionary basis to assist with some of the manual coding of the dictionaries above. These dictionaries include Whissell's Dictionary of Affect in Language (DAL) (Whissell 1989), which labels words along a scale of pleasantness (n=8743); the Affective Norms for English Words (ANEW) word list mean score for manually-code valence (Bradley and Lang 1999) (n=1034); and Turney and Littman's (2003) word list generated using point-wise mutual information for seed positive and negative valence words over a large corpus.

Once merged, the number of positive and negative classifications (in the various dictionaries) for each word was calculated and the highest valence score retained.

Ambiguous and neutral words were dropped – that is, those with valence scores that were equal (or tied) and those classified by any of the dictionaries as "neutral" or "ambiguous." The final merged list includes 33,124 words scored for clearly positive or negative tone. A subset of 3971 words was generated by dropping words that are coded for tone by a single dictionary. The performance of this much smaller subset is nearly identical to the full-version.[12]

It is this 3971-word dictionary that is used for the results that follow.[13] Using the proximity-based procedure described above, the result is a continuous measure of the polarity, or tone, for each party leader and/or party, for each article. A score of zero reflects perfect neutrality; positive scores indicate increasingly positive coverage; and negative scores indicate increasingly negative coverage.

Analysis

How does this automated measure of tone compare with the manually-coded measure? How well does it predict vote intentions in the 2006 election? Recall that the automated sample is not quite the same as the manual sample, so we cannot compare articles directly. We can compare the general trends produced by the two content analyses, however, and do so in Figure 3.

[Figure 3 about here]

There clearly are some similarities, but differences as well.[14] Indeed, there are two differences in particular. First, the automated system produces markedly low estimates of Conservative tone over weeks 5 and 6 of the campaign. We believe this error has to do with the difficulty in attributing tone to one party or the other. When two parties (or leaders) are mentioned in the same sentence, our system necessarily

---

[12] This is likely because many of the dropped words are obsolete, found exclusively in the Roget Thesaurus. It also reflects the infrequent common use of words appearing in one dictionary only, which therefore scarcely affect the overall tone in the automation process. Trials were also conducted using a subset of 6380 subjective words, noted in the literature to improve sentiment analysis (Wiebe 2000). However, this version performed worse than the subset of 3971 used in this study.

[13] By way of example, some of the words with positive valence are: beaming, charity, cognizant, comprehend, credible, curious, dignify, dominance, ecstatic, friend, gain, gentle, justifiably, look up to, meticulous, of_note, peace, politeness, reliability, and success; words with negative valence include: admonish, appall, disturbed, fight, flop, grouch, huffish, hypocritical, impurity, irritating, limp, omission, oversight, rancor, relapse, sap, serpent, untimely, worrying, and yawn.

[14] We provide only a very brief comparison of the automated and manual results here, but recognize that there is much more to do. Towards that end, we are comparing automated results with a random sample of 500 articles, drawn from the automated dataset, and coded by expert coders. The degree to which we expect automated results to be reliable at the level of individual articles is, however, another matter as well. Most work in computational linguistics assumes there will be a good degree of error in each sentence or article, but that the error will be randomly distributed, and thus, given a large enough sample the estimated mean will be correct. It may be, then, that daily averages are very accurate even as results for individual articles are less so. But this is a topic for another paper.

attributes the positive and negative words in that sentence to both parties. This becomes particularly problematic when many sentences mention both parties (and few sentences mention just one party). It seems to matter most in weeks 5 and 6 of this campaign, when there is a considerable volume of horserace coverage noting the degree to which Liberals continue to lose ground to Conversatives. A good number of sentences should be negative for Liberals, but not for Conservatives, and the mistaken attribution pulls Conservative net tone downwards. The other major difference is the decidedly high estimate of Liberal tone in week 2. We have no sense for why this is the case; finding the source of this mis-estimate will require a more careful analysis of the week 2 articles.

[Table 3 about here]

Even so, we are struck by the degree to which automated tone tracks the manual measure, in spite of using a relatively simple bag-of-words, proximity-based approach. We are equally struck by the strength of predictions using this automated measure, presented in Table 3. The MAE for the models using these media measures alone (columns 2 and 5) is roughly 2 – greater than for the manual measure used in Table 1, but well within what we regard as a reasonable range. Conservative and Liberal tone measures are correctly signed in all cases except column 3, where Conservative tone is negative (but also half the size of it standard error). Perhaps most importantly, prediction accuracy is in both cases strengthened by the inclusion of the media measures, albeit only marginally – by .1 percentage points for the Conservatives, and by .2 percentage points for the Liberals.

[Figure 4 about here]

Predictions based on automated media content alone are illustrated in Figure 4. Manual predictions are included in the figure as well, to facilitate a comparison of the two approaches. Resulting are remarkably similar, in spite of different methods and slightly different samples. This similarity speaks to the potential for automated media content analyses in election analyses, and in political science more generally.

Conclusions

This preliminary work has pointed to the potential for both manual and automated media content analysis in campaign-period vote predictions. We have found a good degree of predictive power using relatively simple models of lagged media content (and a series of dummies for pollster effects). The performance of these models is rather impressive, we believe, particularly given that they include only pollster effects and media content measures. And importantly, even when a lagged dependent variable is included, media content improves the accuracy of predictions.

Where the automated analysis is concerned, we do see some opportunities for improvement. On the one hand, we are pleased with the performance of the current system. Most existing work using automated analysis to attribute tone does so only for entire stories; we have attempted and partly succeeded to do so for subjects within stories. On the other hand, we run into difficulties when two subjects (such as

Harper *and* Martin) are mentioned in the same sentence. In these cases we cannot reliably attribute positive or negative language to the right subject. This is a tractable problem, using natural language processing techniques that can distinguish between, for instance, the subject and object of a sentence. This will be a next step in our ongoing work on automated content analysis.

Even with what we see as minor flaws in the automated analyses, preceding results make clear that the general tone of major newspapers in Canada precedes shifts in vote intentions. Whether this is a media *effect* is perhaps not clear. The argument against media effects is that, rather than lead opinion, media content happens to capture and arrange in a readily quantifiable form the evolving mood of the campaign. Journalists are highly attentive to both the campaign and their audience. They react quickly, and measurably, to shifts in the mood of the campaign. They also likely react quickly to shifts in public opinion – not just public opinion generally, but to "opinion leaders."[15] Media content may as a consequence not lead so much as mirror, albeit mirror very efficiently. The public opinion captured in polls is in contrast rather slow and clumsy. So media content leads opinion, but perhaps only in a statistical sense.

The argument for media effects is nevertheless rather strong. We begin with a basic fact: most information about the campaign that citizens receive comes from mass media; it follows that almost all movement over a campaign is a media effect.[16] This is a relatively broad definition of media effects, admittedly. It does not distinguish between mass media acting simply as a conduit for information coming from parties and mass media playing a more active role in selecting and defining the campaign. But note that the kind of information we are extracting from news stories – sentiment-laden vocabulary – is likely to capture the part of media content relating to description and interpretation. That our media measure is likely to capture evaluative language may make more likely the possibility that the media-opinion link discovered here is indeed a causal one.

We clearly lean towards the media effects story, then, though we cannot entirely refute the possibility that media simply reflect evolving trends. For now, knowing that there is a strong connection between the tone of media content and vote intentions may have to suffice. Note, however, that the strength of that connection is great enough that vote predictions based on manually-coded data are clearly possible. And though some flaws still exist, the strength of the automated system used here suggests to us that we are not far away from being able to predict movement in vote intentions using little more than a laptop and access to a full-text news index.

---

[15] On "opinion leaders," see work on the two-step flow in political communications, esp. Lazarsfeld et al. 1944 and Lazarsfeld and Katz 1955.

[16] Assuming the movement is not just random, of course; that is, assuming that movement has something to do with the campaign.

Bibliography

Andrew, Blake. 2007. "Above the Fold and Behind the Link: A Comparative Analysis of Election Shortcuts in Newspapers and on the Web." Paper presented at the 4th European Consortium for Political Research (ECPR) General Conference. Pisa, Italy

Behr, Roy L. and Shanto Iyengar. 1985. "Television News, Real-World Cues, and Changes in the Public Agenda." *Public Opinion Quarterly* 49(1): 38-57.

Blais, Andre, and M. Martin Boyer. 1996. "Assessing the Impact of Televised Debates: The Case of the 1988 Canadian Election." *British Journal of Political Science* 26(2): 143-164.

Bradley, M.M., & Lang, P.J. 1999. *Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings*. Technical report C-1, Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

Brians, Craig Leonard and Martin P. Wattenberg. 1996. "Campaign Issue Knowledge and Salience: Comparing Reception from TV Commercials, TV News and Newspapers." *American Journal of Political Science* 40(1): 172-193.

Craig, Richard. 2000. "Expectations and Elections: How Television Defines Campaign News." *Critical Studies in Mass Communication* 17(1): 28-44.

Druckman, James N. 2004. "Priming the Vote: Campaign Effects in a U.S. Senate Election." *Political Psychology* 25(4): 577-59.

Fan, David P. and Lois Norem. 1992. "The Media and the Fate of the Medicare Catastrophic Extension Act." *Journal of Health Politics Policy and Law* 17(1):37-70.

Fleiss, JL, B. Levin, and MC Park (2003). Statistical Methods for Rates and Proportions (3rd ed), Chichester: Wiley

Fletcher, Frederick. 1981. "Playing the Game: The Mass Media in the 1979 Campaign." In *Canada at the Polls, 1979 and 1980: A Study of the General Elections*, ed. Howard R. Penniman. Washington, D.C.: American Enterprise Institute for Public Policy Research.

————. 1991. *Reporting the Campaign: Election Coverage in Canada*. Toronto: Dundurn Press.

Graber, Doris A. 1976. "Press and TV as Opinion Resources in Presidential Campaigns." *Public Opinion Quarterly* 40(3): 285-303.

Hall-Jamieson, Kathleen and Joseph N. Cappella. 1998. "The Role of the Press in the Health Care Reform Debate of 1993-1994." Pp. 110-31 in Doris A. Graber, Dennis McQuail and Pippa Norris, eds., *The Politics of News: The News of Politics* (Washington DC: CQ Press).

Iyengar, Shanto Iyengar. 1996. "Framing Responsibility for Political Issues." *Annals of the American Academy of Political and Social Science* 546:59-70.

Jackman, Simon (2005). "Pooling the Polls Over an Election Campaign" in Australian Journal of Political Science 40(4): 499-517.

Jamieson, Kathleen Hall. 1992. *Dirty Politics: Deception, Distraction, and Democracy*. New York: Oxford University Press.

Krueger, James S. and Michael Lewis-Beck. 2005. "The Place of Prediction in Politics." Paper presented at the Annual Meeting of the American Political Science Association, Washington D.C.

Krosnick, Jon A., and Donald R. Kinder. 1990. "Altering the Foundations of Support for the President Through Priming." *The American Political Science Review* 84(2): 497-512.

Martindale, C. 1975. *Romantic Progression: The psychology of Literary History*. Washington, D.C.: Hemisphere.

Martindale, C. 1990. The Clockwork Muse: The Predictability of Artistic Change. New York: Basic Books.

McCombs, Maxwell E. and Donald L. Shaw. 1972. "The Agenda-Setting Function of the Mass Media." *Public Opinion Quarterly* 36(2): 176-187.

McDermott Monika L., and Kathleen A. Frankovic (2003). "Horserace Polling and the Survey Method Effects: An Analysis of the 2000 Campaign" in Public Opinion Quarterly, 67(2): 244-264.

Mendelsohn, Matthew. 1996. "The Media and Interpersonal Communications: The Priming of Issues, Leaders, and Party Identification." *The Journal of Politics* 58(1): 112-125.

————. 1994. "The Media's Persuasive Effects: The Priming of Leadership in the 1988 Canadian Election." *Canadian Journal of Political Science* 27(1): 81-97.

————. 1993. "Television's Frames in the 1988 Canadian Election." *Canadian Journal of Communication* 18(2): 149-171.

Mendelsohn, Matthew and Richard Nadeau. 1997. "The Religious Cleavage and the Media in Canada." *Canadian Journal of Political Science* 30(1): 129-146.

Mergenthaler, E. 1996. Emotion-abstraction Patterns in Verbatim Protocols: A New way of Describing Psychotherapeutic Processes. *Journal of Consulting and Clinical Psychology* 64(6): 1306-1315.

Mergenthaler, E. 2008. Resonating Minds: A School-independent Theoretical Conception and its Empirical Application to Psychotherapeutic Processes. *Psychotherapy Research* 18(2): 109-126.

Mullen, A. and Collier, N. 2004. "Incorporating Topic Information into Sentiment Analysis

Models." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain.

Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, Philadelphia, PA.

Patterson, Thomas E. 1993. *Out of Order*. New York: Knopf.

Pennebaker, James W., Martha E. Francis and Roger J. Booth. 2001. Linguistic Inquiry and Word Count: LIWC 2001. Mahway, NJ: Erlbaum Publishers.

Pickup, Mark, and Richard Johnston (2007) "Campaign Trial Heats as Electoral Information: Evidence from the 2004 and 2006 Canadian Federal Election" in Electoral Studies 26:460-476.

Riffe, Daniel, Stephen Lacy and Frederick Fico. 2005. Analyzing Media Messages: Using Quantitative Content Analysis in Research. 2nd Edition. Mahwah, N.J.: Lawrence Erlbaum.

Roget, Peter Mark. 1911. *Roget's Thesaurus of English Words and Phrases*, supplemented electronic version (June 1991). Project Gutenberg Library Archive Foundation.

Soroka, Stuart. 2003. "Media, Public Opinion, and Foreign Policy," *Harvard International Journal of Press and Politics* 8(1): 27-48.

————. 2002. Agenda-Setting Dynamics in Canada. Vancouver BC: University of British Columbia Press.

Soroka, Stuart and Blake Andrew. N.d. "Media Coverage of Canadian Elections: Horserace Coverage and Negativity in Election Campaigns," in Linda Trimble and Shannon Sampert, eds., *Mediating Canadian Politics* (Pearson, forthcoming).

Stone, Deborah. 1989. "Causal Stories and the Formation of Policy Agendas," *Political Science Quarterly* 104(2):281-300.

Stone, P.J., D.C. Dumphy and D.M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. M.I.T. Press: Cambridge, MA.

Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet-Affect: An Affective Extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, IT.

Turney, P and M. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)* 21(4).

Whissell, C. 1989. The Dictionary of Affect in Language. In R. Plutchnik and H. Kellerman, eds. *Emotion: Theory and Research*. New York, Harcourt Brace, 113-131.

Wiebe, Janyce M. 2000. *Learning Subjective Adjectives from Corpora*. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*, Austin, TX.

Wilson, R. Jeremy. 1980. "Media Coverage of Canadian Election Campaigns: Horserace Journalism and the Meta-Campaign." *Journal of Canadian Studies* 15(4): 56-68.

Wagenberg, R. H., W. C. Soderlund, W. I. Romanow, E. D. Briggs. 1988. Campaigns, Images and Polls: Mass Media Coverage of the 1984 Canadian Election. *Canadian Journal of Political Science* 21(1): 117-129.

Weaver, David H. 1996. "What Voters Learn from Media." *Annals of the American Academy of Political and Social Science* 546: 34-47.

Table 1. 2006 prediction models, manually-coded media data

	DV: 2006 Vote Intentions $_t$					
	Conservative			Liberal		
	1	2	3	4	5	6
$\sum$ **CPC** $_{t-4,5,6}$		**35.750****	**7.760**		**-30.854****	**-18.169****
		(16.217)	(7.747)		(14.620)	(4.000)
$\sum$ **Liberal** $_{t-4,5,6}$		**-22.280**	**7.110**		**38.487****	**7.441**
		(20.094)	(9.435)		(18.116)	(5.161)
DV $_{t-4}$	.971**		.977**	.970**		.922**
	(.069)		(.094)	(.077)		(.050)
Pollster Effects						
SES	-.241	-.561	-.521	.819	.654	.647*
	(.480)	(1.448)	(.649)	(.511)	(1.305)	(.352)
SC	-.191	1.341	.027	-.135	-1.783**	-.581**
	(.247)	(.817)	(.387)	(.278)	(.737)	(.209)
Leger	-1.003**	-2.391*	-1.083*	-.254	1.938*	.106
	(.469)	(1.237)	(.568)	(.518)	(1.116)	(.316)
Decima	.325	2.448	-.414	-1.474**	-2.962**	-1.023**
	(.466)	(1.559)	(.751)	(.492)	(1.405)	(.393)
Ipsos	-.193	.666	-.462	.482	-1.219	.340
	(.441)	(1.274)	(.581)	(.477)	(1.149)	(.321)
Environics	-1.577	4.773	-1.913	.492	-7.165**	-.899
	(1.204)	(3.396)	(1.651)	(1.300)	(3.061)	(.891)
Pollara	.154	-2.140	.472	.049	2.205	-.761*
	(.653)	(1.661)	(.785)	(.698)	(1.498)	(.434)
Constant	3.255	29.192**	4.499	-1.457	40.949**	2.848
	(2.623)	(4.776)	(3.194)	(3.144)	(4.306)	(2.355)
Rsq	.895	.564	.916	.892	.682	.978
N	47	47	47	47	47	47
Accuracy						
MAE	.952	1.850	.866	1.013	1.665	.468

Cells contain OLS coefficients with standard errors in parentheses.
* $p < .10$; ** $p < .05$.

Table 2. 2004 prediction models, manually-coded media data

	DV: 2004 Vote Intentions $_t$					
	Conservative			Liberal		
	1	2	3	4	5	6
$\sum$ CPC $_{t-4,5,6}$		**13.513**	**41.043****		**-9.125**	**-14.979**
		(9.570)	(7.432)		(7.830)	(9.610)
$\sum$ Liberal $_{t-4,5,6}$		**.046**	**4.775**		**1.126**	**.030**
		(10.531)	(6.257)		(8.616)	(8.988)
DV $_{t-4}$	.393**		.922**	.463**		.380
	(.113)		(.165)	(.218)		(.316)
Pollster Effects						
SES	.499**	1.260**	.550*	-.220	-1.144**	-.790*
	(.221)	(.344)	(.266)	(.321)	(.282)	(.416)
SC	.000	.000	.000	.000	.000	.000
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
Leger	-.351	-.708	-1.364**	-.025	-.197	-.159
	(.439)	(.908)	(.556)	(.476)	(.743)	(.775)
Decima	.000	.000	.000	.000	.000	.000
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
Ipsos	-1.790**	-.764	-1.228**	-.221	-.833*	-.951*
	(.357)	(.535)	(.338)	(.372)	(.438)	(.485)
Environics	.000	.000	.000	.000	.000	.000
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
Pollara	.000	.000	.000	.000	.000	.000
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
Constant	20.153**	31.744**	6.845	18.246**	34.940**	21.310*
	(3.397)	(2.341)	(4.651)	(7.513)	(1.915)	(11.483)
Rsq	.679	.728	.918	.462	.769	.751
N	31	31	31	31	31	31
Accuracy						
MAE	.629	.610	.396	.828	.545	.524

Cells contain OLS coefficients with standard errors in parentheses.
* $p < .10$; ** $p < .05$.

Table 3. 2006 prediction models, automated media data

	DV: 2004 Vote Intentions $_t$					
	Conservative			Liberal		
	1	2	3	4	5	6
$\sum$ **CPC** $_{t-4,5,6}$		**1.214**	**-.314**		**-1.007**	**-.134**
		(1.543)	(.580)		(1.491)	(.694)
$\sum$ **Liberal** $_{t-4,5,6}$		**-1.162**	**-.811**		**1.496**	**1.326**
		(1.449)	(.536)		(1.399)	(.648)
DV $_{t-4}$	1.004**		.976**	.983**		.942**
	(.059)		(.068)	(.082)		(.087)
Pollster Effects						
SES	-.044	-.626	-.071	.855	.885	.892
	(.434)	(1.380)	(.512)	(.544)	(1.333)	(.617)
SC	-.297	.778	-.563	.022	-1.389	.192
	(.223)	(.941)	(.360)	(.292)	(.909)	(.445)
Leger	-.777*	-2.758*	-1.079*	-.051	2.093	.744
	(.423)	(1.430)	(.541)	(.537)	(1.381)	(.651)
Decima	.107	1.248	-.171	-1.386**	-1.897	-.994
	(.424)	(1.452)	(.546)	(.528)	(1.403)	(.654)
Ipsos	-.283	.947	-.080	.037	-1.042	-.034
	(.410)	(1.204)	(.451)	(.513)	(1.163)	(.546)
Environics	-1.429	3.042	-1.243	.241	-4.463	-.551
	(1.155)	(3.496)	(1.326)	(1.460)	(3.377)	(1.603)
Pollara	.308	-3.045	.610	-.371	3.342*	-1.106
	(.599)	(1.862)	(.734)	(.770)	(1.798)	(.927)
Constant	1.777	33.792**	3.446	-1.933	33.810**	-1.210
	(2.220)	(3.858)	(2.561)	(3.421)	(3.727)	(3.650)
Rsq	.902	.441	.926	.853	.499	.896
N	52	52	52	52	52	52
Accuracy						
MAE	.935	2.114	.839	1.155	2.023	.917

Cells contain OLS coefficients with standard errors in parentheses.
* $p < .10$; ** $p < .05$.

Figure 1. 2006 predictions, media content only, manually-coded media data

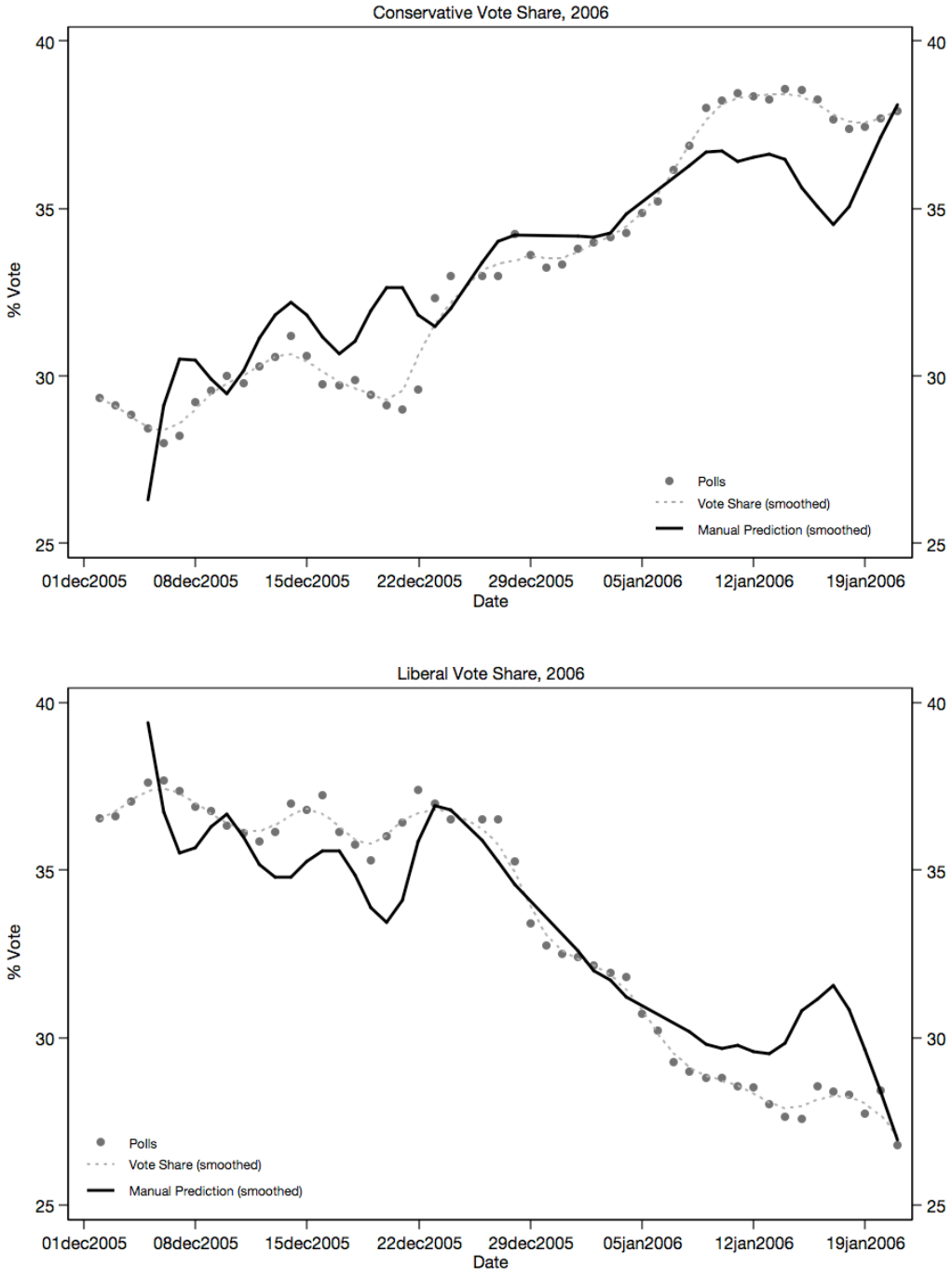

Figure 2. 2004 predictions, media content only, manually-coded media data

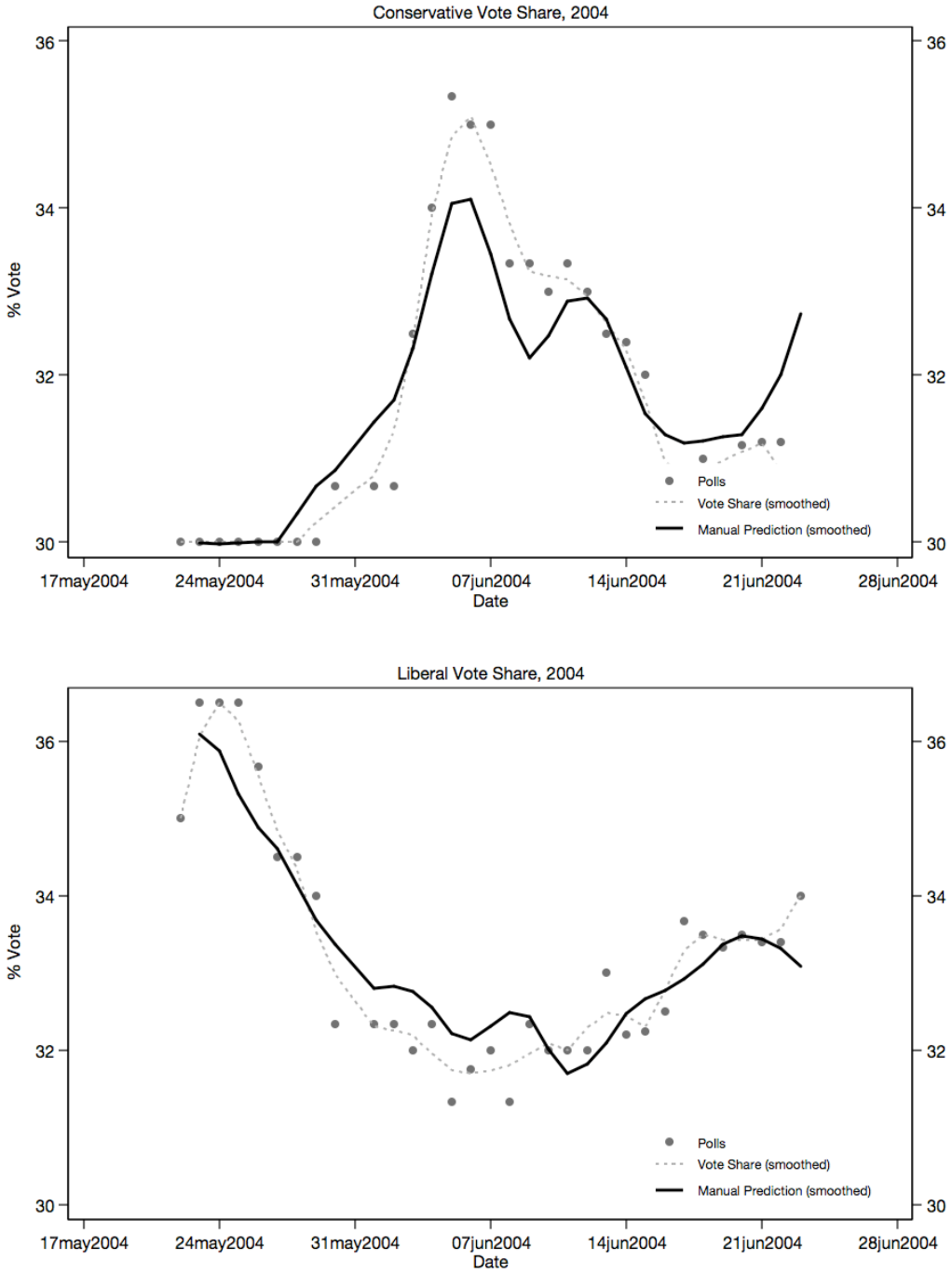

Figure 3. Manual and Automated 'Net Tone'

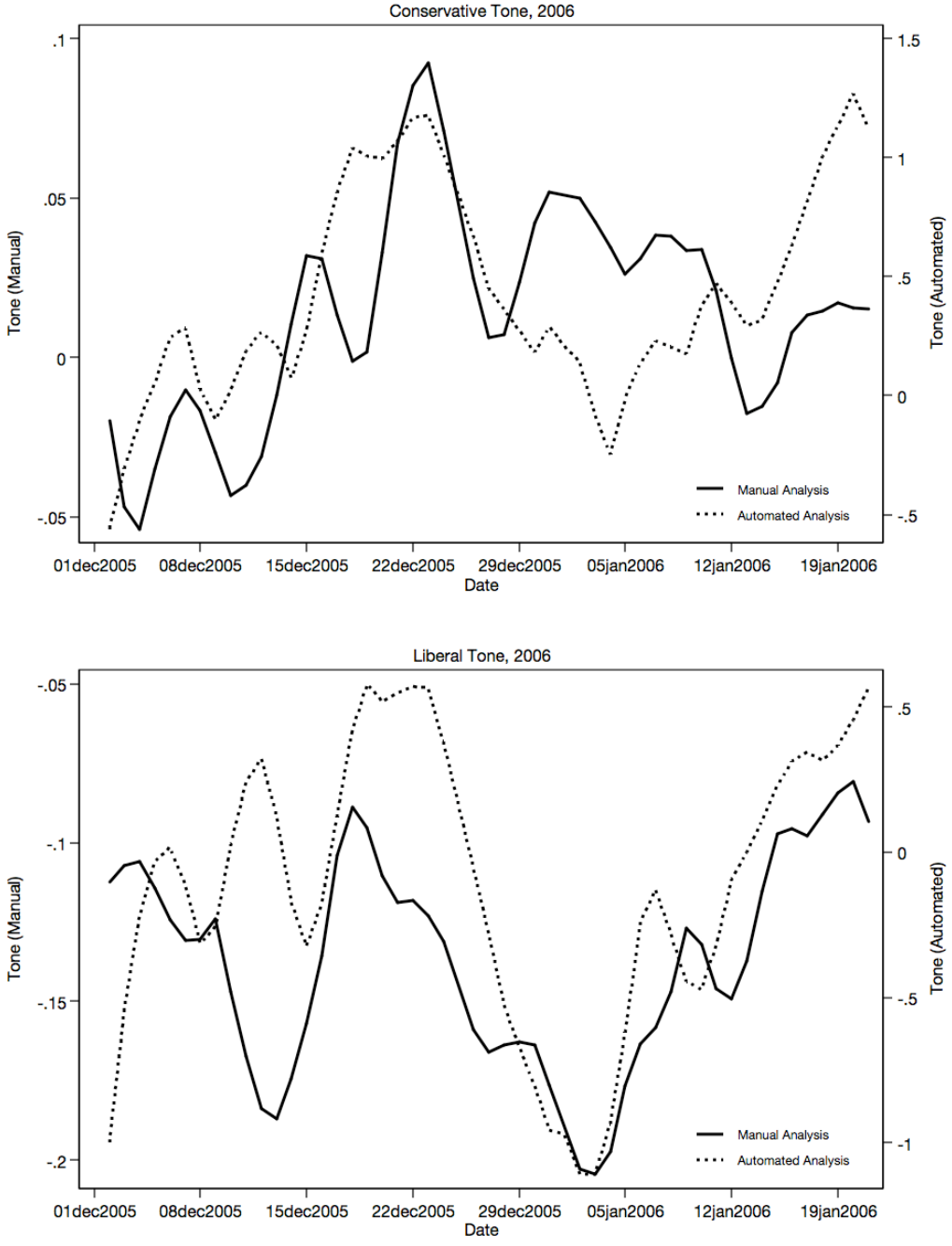

Figure 4. 2006 predictions, media content only, automated media data