Political Communication, 29:205–231, 2012

ISSN: 1058-4609 print / 1091-7675 online
DOI: 10.1080/10584609.2012.671234

Routledge
Taylor & Francis Group

Affective News: The Automated Coding of Sentiment in Political Texts

LORI YOUNG and STUART SOROKA

An increasing number of studies in political communication focus on the "sentiment" or "tone" of news content, political speeches, or advertisements. This growing interest in measuring sentiment coincides with a dramatic increase in the volume of digitized information. Computer automation has a great deal of potential in this new media environment. The objective here is to outline and validate a new automated measurement instrument for sentiment analysis in political texts. Our instrument uses a dictionary-based approach consisting of a simple word count of the frequency of keywords in a text from a predefined dictionary. The design of the freely available Lexicoder Sentiment Dictionary (LSD) is discussed in detail here. The dictionary is tested against a body of human-coded news content, and the resulting codes are also compared to results from nine existing content-analytic dictionaries. Analyses suggest that the LSD produces results that are more systematically related to human coding than are results based on the other available dictionaries. The LSD is thus a useful starting point for a revived discussion about dictionary construction and validation in sentiment analysis for political communication.

Keywords content analysis, media tone, methodology

Political discourse cannot be reduced to mere factual information—the tone of a text may be as influential as its substantive content. Indeed, numerous studies have focused on the tone or sentiment of news content, political speeches, and advertisements.[1] Moreover, a substantial and growing body of research suggests that affect[2] is a central component of individual decision making and political judgment generally, as well as the processing of media information in particular.[3] Negative affect appears to be particularly prominent in the human psyche, and in politics as well.[4] The reliable and valid analysis of sentiment is, in short, a critical component of a burgeoning field of research in political communication, and political science more broadly.

The growing interest in, and importance of, measuring sentiment coincides with a dramatic increase in the volume of digitized information. Computer automation has a great deal of potential in this new media environment. Automation is very efficient and

Lori Young is a doctoral candidate at the Annenberg School for Communication, University of Pennsylvania. Stuart Soroka is Associate Professor and William Dawson Scholar in the Department of Political Science, McGill University.

The authors are grateful to Mark Daku, who programmed Lexicoder; to Marc André Bodet and Blake Andrew for their work using the LSD in its early stages; to Christopher Wlezien and Robert Erikson, for providing us some U.S. polling data with which to further test the dictionary; and to the editor and anonymous reviewers, whose comments were critical to this final version of the article.

Address correspondence to Stuart Soroka, Department of Political Science, McGill University, 855 Sherbrooke St. West, Montreal, QC, H3A 2T7 Canada. E-mail: stuart.soroka@mcgill.ca

becoming easier to implement as new software is developed and lexical resources become more widely available. Our objective here, then, is to outline and validate a new automated measurement instrument for sentiment analysis in political texts. Our instrument uses a dictionary-based approach consisting of a simple word count of the frequency of keywords in a text from a predefined dictionary. There are a number of machine-readable sentiment lexicons currently available for automation. However, each has been compiled for specific types of research in various disciplines using diverse methodologies. Consequently, they vary widely with respect to sentiment categories, coding schemes, and scope of coverage. Indeed, there are to our knowledge no comparative studies of existing lexicons used in sentiment analysis in political communication. Moreover, we find that proprietary restrictions occasionally impinge on the assessment and use of such resources, raising concerns about replicability and development in the field. Many lexicons are also temporally or corporally specific. Our goal is to develop a sentiment dictionary that is more broadly applicable, across a wide range of research foci in political communication.

We do so by combining and standardizing three of the largest and most widely used lexical resources to create a comprehensive valence dictionary of positive and negative words. The scope and performance of the resulting dictionary is then compared to six other commonly used sentiment lexicons, as well as to the three from which it was composed. We automate the analysis of positive and negative tone in *New York Times* coverage across four topics: economy, environment, crime, and international affairs. We compare results not just across sentiment dictionaries, but more importantly with results from trained human coders. Results suggest that the dictionary developed here, the Lexicoder Sentiment Dictionary (LSD), performs somewhat better than others. This, combined with the fact that the LSD is freely available and easily adaptable, makes the dictionary proposed here a valuable step forward in automated content analysis of political communication.

The following sections outline the design and reliability of the dictionary in some detail. A final section then provides one example of how the dictionary can be used, by extending previous work on the relationship between campaign-period vote intentions and the tone of media content. We extend previous work on the 2006 Canadian election campaign, comparing directly the previous manually coded results with those using the LSD; results demonstrate the strength of the dictionary and also speak to the role of media in election campaigns. First, however, the next sections review the relevant literature concerning media effects, describe in some detail the state of research in automated content-analytic techniques as they pertain to current work in political communication, and consider some of the advantages and limitations of a dictionary-based approach.

Media Affect

The development of a dictionary for automated content analysis below stems first and foremost from our interest in media effects, and more generally the role of media in representative democracy. It is axiomatic that, across all modern representative democracies, mass media play a central role in everyday politics. Media both reflect and inform public opinion; many of us are dependent on mass media for much of the information we require in order to be effective democratic citizens. Scholars are thus perennially concerned with the content of mass media, as well as the potential consequences that content may have on political judgment and behavior. There are vast bodies of work detailing the many ways in which media can affect public preferences on political issues. Much of this work has been interested in large-scale content analysis of media content.

Most relevant to our work here is research focused on capturing the tone of media content. This body of research is wide and varied—it includes, for instance, work on the tone of news coverage of presidents and political parties (e.g., Eshbaugh-Soha, 2010; Farnsworth & Lichter, 2010; Ottati, Steenbergen, & Riggle, 1992; Soroka, Bodet, Young, & Andrew, 2009), research dealing with the tone of economic news coverage (e.g., Gentzkow & Shapiro, 2010; Lowry, 2008; Nadeau, Niemi, Fan, & Amato, 1999; Soroka, 2006), and work reflecting other diverse interests, such as Cho et al.'s (2003) research on the "emotionality" of television and print media coverage of the 9/11 terrorism attacks.

The affective content of news is also related to the body of literature focused on symbolic language and issue frames. This work views modern politics largely as a struggle over language, a battle to define terms and frame the debate (see, e.g., Edelman, 1985; Hart, 2000b). It suggests that symbolic language and framing influences the way people think about particular issues (see, e.g., Iyengar, 1996; Quattrone & Tversky, 1988). Shifting frames can change the affective composition of the media, creating narratives that construct and shape perceptions of social and political and reality (e.g., Johnson-Cartée, 2005; Shenhav, 2006). Driven by dramatic, novel, and negative information, affective narratives can inform judgments, sometimes quite independent of real-world events (e.g., McComas & Shanahan, 1999). And while the automated identification of frames is not our focus here, we view sentiment as one central component in the empirical study of issue frames.

In short, a wide range of work suggests that the "tone" or "sentiment" of text matters to our understanding of both the content and effects of mass media in modern representative democracy. It matters for our understanding of media content, political behavior, and policy-making. Given the widespread interest in, and demonstrated importance of, tone in media content, coupled with the increasing use of computer automation, we believe that it is important to consider the extent to which tone can be captured reliably and validly using automated systems.

Computer Automation

Computer automation has become a mainstay of empirical research in the study of political communication. Since the 1950s scholars have been developing computer-assisted methods to analyze textual information in new and interesting ways. As mass media have expanded, so too has the volume of political text, and this text is increasingly readily available electronically. Increasing volumes of digital information have been met with increasingly sophisticated content-analytic methodologies. Many of these are computer automated. Automation can facilitate the analysis of enormous bodies of data in meaningful ways, where labor-intensive manual content analyses often fall short, either because of time and budgetary constraints or because of difficulties obtaining intercoder reliability.

Broadly speaking, automated content analysis can be undertaken as a statistical or a nonstatistical endeavor. Machine-learning techniques using statistical classification exploded with computational advances in the 1990s. This method does not rely on predefined dictionaries; rather, data are generated from the text itself using statistical classifiers. Supervised machine learning (SML) involves identifying various features (prevalent words or word patterns) in a set of "reference" texts with a known a priori class or category. Reference texts are manually classified or selected to be the best possible representation of the category to be coded. Much rests, then, on the quality and representativeness of the reference texts, which are used to "train" classifiers to recognize or predict the class of unknown texts according to the presence of linguistic features "learned" from the reference texts.[5]

Unsupervised machine learning (UML) does not rely on reference texts or predefined categories. Based on latent semantic analysis, the method uses matrix algebra to measure word associations (i.e., word clusters, local co-occurrence, pairwise patterns) within and between texts to infer unknown categories, much like factor analysis (Hogenraad, McKenzie, & Péladeau, 2003; Landauer & Dumais, 1997). It is relatively easy to implement and does not require extensive pre-coding or a priori dictionaries. SML (as well as the dictionary-based approach discussed below) relies on predefined classification schemes that map onto the text, giving meaning to the words; in contrast, UML is used to discover new or unknown categories inductively, using the correlation of words to give meaning to a text.[6]

The nonstatistical dictionary-based approach (also referred to as frequency or categorical analysis) is, in terms of implementation, much simpler: It involves counting the frequency of definitive keywords in a text. A key feature of this approach is use of a machine-readable dictionary of a priori categories. Herein lies the challenge: A good dictionary, particularly for something like sentiment, is very difficult to develop. Nevertheless, numerous content-analytic dictionaries have been developed for automated analysis featuring a range of topic and sentiment lexicons.

With a well-defined and comprehensive dictionary, a basic word count can provide a powerful and reliable analysis of the topical and affective composition of a text. Existing applications range from analysis of children's writing to discern their affect toward police (Bolasco & Ratta-Rinaldi, 2004) to predicting the variance of firm account earnings and stock returns by counting negative words near the keyword "earn" in economic journals (Tetlock, Saar-Tsechansky, & Macskassy, 2007). The approach has been used to analyze political communications since the 1960s. Stone and colleagues (Stone, Bales, Namenwirth, & Ogilvie, 1962; Stone, Dumphy, & Ogilvie, 1966) first used the General Inquirer (GI) dictionary to compare the tone of political speeches by various candidates. Hart (1984, 2000a) has used a similar approach since the 1980s to analyze president rhetoric and campaign style. Hart's DICTION program has become a mainstay in content-analytic work on political rhetoric and discourse and has been used in over 50 studies to analyze political speech, differentiate news by genre, study religious ideology, analyze corporate publications, and so forth.[7]

Advantages and Challenges in Automated Analysis

There are some clear advantages to using computer automation for text analysis, including efficiency, scope, and reliability. Dictionary-based results have the additional advantage of parsimony. Constructing a dictionary is quite an investment, to be sure. Once constructed, however, computation is very easy to implement. Certain dictionaries may be better suited for some texts than others, of course, but in each case we know exactly what is being applied, and all cases are thus directly comparable. They are also perfectly reliable, in the sense that they produce exactly the same results at the article level, whether one analyzes 10 articles or 10 years of news.

This increase in reliability does not necessarily reflect greater validity, of course. Automation is typically capable of lexical and syntactic analysis and certain types of semantic or discourse analysis (see further discussion below). But some types of textual analysis are much less well suited for automation, dictionary-based or otherwise. Automation counts but does not rate entries; it identifies but does not interpret semantic patterns; it quantifies concepts but not symbols. To be clear: We readily acknowledge that

there are many questions of interest in content analysis that are still beyond the capability of computers.

That said, automated analysis clearly does have value in particular contexts. In one sense, it is simply a different level of analysis. Borrowing Hart's (2001) analogy, manual coding may be likened to the perspective of a beat cop in a specific neighborhood, rich in context and detail-oriented, while computer automation offers a bird's eye view, like a helicopter pilot circling the city to monitor overall crime patterns. The methods are complementary—each perspective generates useful information the other cannot see.

We should also not regard computer automation as somehow less refined than manual forms of textual analysis. Automation may be especially well-suited for certain questions that would be difficult for humans to code. For instance, human coders may be limited in their ability to identify certain types of latent (rather than manifest) content. And as we will see below, many dictionaries are constructed to capture complicated latent cognitive and affective concepts, such as the degree of primordial versus conceptual thought (Martindale, 1975, 1990), or "certainty" operationalized by uses of the verb "to be" (Hart, 1984). Indeed, the psycholinguistic underpinnings of many content-analytic dictionaries point to a level of sophistication in computer automation not always noted by its critics.

That said, there are least two main challenges to automated techniques, each of which deserves mention here. First, most automated systems process words regardless of order or context using the so-called "bag-of-words" approach; that is, they assume "semantic independence." The strategy is "based on the assumption that the words people use convey psychological information over and above their literal meaning and independent of their semantic context" (Pennebaker, Mehl, & Niederhoffer, 2003, p. 550). Obviously, this assumption does not always hold. For instance, many scholars have noted that tone is far more dependent on the relation between words than topic (Murphy, Bowler, Burgess, & Johnson, 2006; Pang, Lee, & Vaithyanathan, 2002; Thomas, Pang, & Lee, 2006; Wilson, Wiebe, & Hoffman, 2005). For instance, multiple speakers and/or topics can make the attribution of tone difficult—the tone of coverage about a political actor or topic is not necessarily reflected in the overall tone of an article.[8] Depending on the goals of the research at hand, however, it may not be necessary to fully disentangle the semantic relationship between actors, topics, and opinions. For texts with multiple topics and speakers, unattributed tone at the document level simply reflects the overall tone of a document. (We discuss this in more detail below.)

A tougher challenge to this assumption is the tendency for the linguistic markers of tone to be context specific. Consider how the meaning of the word "happy" changes when it follows the word "not" or the difficulty determining the meaning of homographs such as "right," "lie," or "well." One strategy to mitigate the impact of contextual language is the preprocessing of text in order to standardize words and phrases, disambiguate homographs, and account for basic negation patterns. In our analysis we apply extensive preprocessing, the development of which is described below. In this way we are able, at least modestly, to move beyond a simple bag of words.

The second general concern with automation is the assumption of additivity—that is, every instance of every word contributes isomorphically to the output. In natural language, of course, certain words may carry more weight than others. "Evil" may matter more than "bad," for instance. The point is well taken; however, it is not clear that this is an issue of content analysis so much as one of psychology. Technically speaking, weights are easy to apply. There are numerous examples of automated dictionaries that use weights or apply modifiers to try to overcome the assumption of additivity (e.g., Subasic & Huettner, 2001). What seems to be lacking is a good theory as to how the weights should be applied based

on the varied effects of particular words. This does not negate the fact that the potentially differential weighting of words can present real problems for automated analyses. On the contrary, it is an important reminder that automation is simply a lexical scan of the frequency of words used. And there is, of course, no substitute for careful data interpretation.

Sentiment Lexicons

Since the 1960s, scholars have been developing psycholinguistic lexicons coded for basic affective and cognitive dimensions or tagged for valence to categorize the positive and negative connotations they carry. There are now numerous machine-readable dictionaries available for research, but they vary widely with respect to categories and scope of coverage. They include, for example, the following: from political science, the GI (Stone et al., 1966); from communication, DICTION (Hart, 2000a); from psychology, Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001), the Regressive Imagery Dictionary (RID) (Martindale, 1975, 1990), and TAS/C (Mergenthaler, 1996, 2008); from behavioral science, Affective Norms for English Words (ANEW) (Bradley & Lang, 1999); from literature, Whissell's Dictionary of Affect in Language (DAL) (Whissell, 1989); from linguistics, WordNet-Affect (WNA) (Strapparava & Valitutti, 2004); and from computational linguistics, Turney and Littman's (2003) pointwise mutual (PMI) information wordlist, as well as the ubiquitous Roget's Thesaurus (hereafter Roget's) (Roget, 1911).

These dictionaries have been compiled for a variety of research projects across disciplines. Consequently, there has been a range of methodological approaches to dictionary construction. In some cases dictionaries are compiled from previously generated word lists (e.g., GI)[9]; in others codes are manually attributed by expert coders or panels of judges (e.g., LIWC); in others words are tagged using computer automation based on patterns in natural language (e.g., PMI) or the linguistic properties of words (e.g., WNA); and others still are derived from experimental methods (e.g., ANEW, DAL) or iterative processes combining a number of different approaches (e.g., DICTION, TAS/C). Each method measures something slightly different. Generally speaking, expert coding seeks to capture the definitive meaning of words, automation captures contextual or common usage, and experimental methods capture something closer to perceptions of words.

Construction of a valence lexicon is particularly challenging because the semantic category of valence itself appears to be structurally fuzzy (Andreevskaia & Bergler, 2006; Subasic & Huettner, 2001). The ambiguity of the category poses a challenge to researchers who rely on discrete categories (positive versus negative) for frequency analysis, and efforts to resolve ambiguity tend to further limit the scope of coverage. Two main approaches are adopted to address the ambiguity of valence in dictionary construction: Researchers either disambiguate the dictionary or attempt to score entries according to their centrality to a given category.

For instance, the GI is the oldest and most expansive dictionary of its kind. By consequence, its large valence categories of positive and negative words tend to be overly general and lack discriminative capacity. Few researchers use the dictionary without encountering the need to generate expanded and/or disambiguated versions (e.g., Hogenraad et al., 2003; Kennedy & Inkpen, 2006; Pennebaker et al., 2003; Scharl, Pollach, & Bauer, 2003).[10]

Rather than removing ambiguity, some seek to preserve it by adopting "fuzzy logic" and "continuous valence" categories. Andreevskaia and Bergler (2006) maintain that discrepancies in the dictionaries and inter-annotator disagreements are "not really errors but

a reflection of the natural ambiguity of the words that are located on the periphery of the sentiment category"; disagreement reflects the "structural property of the semantic category" (p. 4). Other work reflects a similar belief; researchers adopting this approach have thus calculated the degree of centrality to a category by (manually or statistically) weighing a word according to all possible meanings, magnitude or intensity, or lexical relations with other words (Andreevskaia & Bergler, 2006; Subasic & Huettner, 2001).[11] Resulting dictionaries may be more precise, but they are often quite limited in their scope. Moreover, they tend to be computationally sophisticated, and only a fraction of such lexicons are available for research, making replication and improvement difficult.

Indeed, notwithstanding concerns about ambiguity, most dictionaries are generated for a particular purpose or genre of text, and as a consequence tend to be temporally and corporally specific. For instance, TAS/C was created to measure emotional tone and abstraction in psychotherapy sessions; DAL was developed to analyze the affective content of poetry and literature; RID was designed to distinguish between primordial and conceptual thinking; and DICTION was developed primarily to understand the rhetoric of speechmakers. Closest to our purposes are GI and LIWC, both of which were developed to analyze various affect categories in political texts.

Indeed, the dictionaries listed above show stunningly little overlap, and where they do overlap codes are often discrepant. (There are only two words that appear in all nine of the dictionaries analyzed here.) To be clear: Despite their varying uses, each dictionary relies on similar underlying constructs relating to various sentiment categories, yet the universe of words and the coding schemes vary greatly. Scholars have yet to construct a universal sentiment lexicon that can be exported across diverse corpora; despite many advances, a definitive lexicon does not exist (Athanaselis et al., 2005; Grefenstette, Qu, Evans, & Shanahan, 2004). The challenge remains, then, to expand the scope of a sentiment dictionary without compromising its accuracy. Below we outline a method to merge, standardize, and disambiguate three of the largest and most widely used lexical resources to create a comprehensive valence dictionary, which we hope meets some of the challenges presented above and which proves broadly applicable for scholarship on the tone of political communication.

The Lexicoder Sentiment Dictionary (LSD)

In a first effort to produce a comprehensive dictionary coded for valence, aimed primarily at news content but potentially useful elsewhere as well, we merge and standardize three widely used and publicly available affective lexical resources from political science, linguistics, and psychology. The resulting Lexicoder Sentiment Dictionary (LSD) is a broad lexicon scored for positive and negative tone and tailored primarily to political texts.

LSD is comprised of words from Roget's Thesaurus, the GI, and the RID.[12] Roget's is the only truly comprehensive word list scored for sentiment. Our goal was to attribute the valence code that a word takes in most contexts. From Roget's, then, we include words from all categories that we identified as positive or negative. This includes, for example, positive categories such as "benevolence," "vindication," "respect," "cheerfulness," and "intelligence" and negative categories such as "insolence," "malevolence," "painfulness," "disappointment," and "neglect" ($n = 47{,}596$).[13] From the GI, we include two broad valence categories labeled "Positiv" and "Negativ" ($n = 4{,}295$). From the RID, we include the positive categories "positive affect" and "glory" and the negative categories "chaos," "aggression," "diffusion," "anxiety," and "sadness" ($n = 1{,}056$).

We first sought to attribute a single code to every word per dictionary.[14] Words found in more positive than negative categories were coded as positive, and vice versa. Ambiguous and neutral words were not included—that is, those found in an equal number of positive and negative categories and those independently coded as neutral or ambiguous by any of the dictionaries. Each word was then classified as positive or negative in the LSD if (a) the word appeared in all three dictionaries and was consistently coded as positive or negative, or (b) the word appeared in just two of the dictionaries but was consistently coded as positive or negative. Additional analysis was required for the remaining words—those mentioned in just one dictionary or those for which there were coding discrepancies across the three dictionaries. Each of these words was considered for inclusion manually. Ambiguously coded words were included if the discrepancy was easy to resolve. Otherwise, contextual analysis (described below) was employed to make final decisions about the remaining words.

Few automated methods attempt to disambiguate homographs,[15] and standard bag-of-words approaches gloss over context entirely. We employed a number of strategies to address these issues, at least minimally. First and foremost, we made liberal use of WordStat's[16] keyword-in-context (KWIC) feature. KWIC functionality is available in many software packages, allowing content analysts to examine different uses of the same word in a corpus. LSD entries were analyzed in context using some 10,000 newspaper articles on a wide range of topics (selected randomly from a database of front-page news stories in major Canadian dailies over a 20-month period), enabling us to identify dictionary entries with multiple word senses and tricky contextual usage. Ambiguous entries were confirmed, dropped, or disambiguated in one of two ways. In some cases ambiguous words were replaced with contextual phrases that capture a particular use or sense of a word. For example, the homograph "lie" is only negative in certain contexts. To capture negative senses only, the dictionary entry "lie" is replaced with several phrases, including "a lie" and "lie to." In other cases disambiguation occurs in the preprocessing phase, to which we turn in a moment.

Contextual analysis was also employed to analyze several problematic word categories. A number of direction words indicating an "increase" or "decrease" initially found their way into the dictionary, even though they do not have a clear tone. For instance, the verb "augment" is listed in many positive categories, though the tone clearly depends on what is being augmented. Likewise, "decline" is often listed as negative, even though it is not so if something negative is declining (e.g., unemployment). A list of economic terms was also analyzed, given the prevalence of economic terms to which tone is attributed. For instance, this process removed words such as "profit" and "credit." Additionally, common stop-words were removed and alternative spellings added. Contextual analysis also facilitated the addition of numerous dictionary entries—many particular to political news reporting—that were not present in any of the core lexicons ($n = 1{,}021$). Finally, all dictionary entries were truncated to capture inflected variations, provided that the inflected forms retained the same tone. In cases where the tone of inflected words differed from the original entry, truncation was not applied; instead, inflected variations were individually added to the dictionary with appropriate codes for tone. The final dictionary comprised 4,567 positive and negative words.[17]

Finally, contextual information was employed to generate several preprocessing modules, which format the text prior to content analysis to facilitate word sense disambiguation and contextual analysis of affective language.[18] The first preprocessing module standardizes and/or removes punctuation. The second removes capitalized words (other than the first word of a sentence). The logic here is to remove proper nouns, which should by

definition not have tone. The third module processes basic negation phrases. Standardizing negation allows for a variety of negated phrases to be captured with a limited number of dictionary entries. The preprocessor first replaces negation words such as "no," "never," "neither," "hardly," "less," and so forth with "not." Second, various verbs and pronouns following negation words are removed—"not going to," "not feeling," "hardly any," "nor were they," "no one was," "without much," and so forth are replaced with "not." Standardized in this manner, the dictionary entry "not good" captures a range of negative phrases including "not at all good," "hardly a good idea," "no one good," "nor was it good," "without goodness," and many more. Finally, the main body of the preprocessor removes "false hits" for dictionary entries, where false hits are topical, multi-toned, or non-tonal instances of a dictionary entry. For example, multi-toned phrases such as "good grief," "losing hope" and "tears of joy" are processed to remove the toned words "good," "hope," and "tears"; non-tonal phrases such as "an awful lot," "crude oil," and "child care" are processed to remove the toned words "awful," "crude," and "care." The preprocessing of over 1,500 words and phrases facilitates basic word sense disambiguation and the contextualization of many commonly used sentiment words and phrases.[19] Some of the most nuanced entries in the dictionary rely on a combination of contextual phrases, truncation, negation, and preprocessing.

Testing: Data and Results

Does the LSD work? More precisely, does the LSD produce codes that are (a) consistent with human coding and (b) more consistent with human coding than other available sentiment dictionaries?

LSD is just the dictionary itself, of course—with some minor reformatting, it can be used by any number of available software packages. We implement it, and all other dictionaries used here, in Lexicoder, which is a freely available java-based, multiplatform software that implements frequency analysis for any number of user-written categorical dictionaries.[20] As the name suggests, it was developed alongside the LSD.

Our aim in this section, then, is to compare the reliability and validity of the LSD to several commonly used lexicons. More specifically, we compare results using the LSD with results using the following[21]:

- LIWC, from which we include the positive category "positive emotion" and the negative categories "negative emotion," "anxiety," "anger," and "sadness" (n = 1,502)
- WNA, from which we use a subset of affective words generated from WordNet synsets labeled "positive" or "negative" (n = 1,640)
- TAS/C's "emotional tone" word list, which is labeled on the dimension pleasure-displeasure (n = 4,058)
- Mean scores from DAL, which labels words along a scale of pleasantness (n = 8,743)
- Mean scores from ANEW, which labels words along a scale of pleasure (n = 1,034)
- Point-wise mutual information scores from the PMI, which are based on the proximity of entries to positive and negative seed words in a text (n = 1,719)

WNA and TAS/C did not require recoding to generate a valence dictionary. In the case of LIWC, we simply collapsed the four negative categories. DAL, ANEW, and PMI posed a challenge, since they are measured on a continuous scale. Here, we had little indication of where the cut-points between positive, neutral, and negative might be. Thus, we divided

each dictionary into terciles based on the scale capturing sentiment, putting the top tercile into the positive category, the bottom tercile into the negative category, and omitting the middle third (which upon inspection was indeed comprised mostly of ambiguous or neutral words).

We also include in our comparison the GI, RID, and Roget's—the three dictionaries from which the LSD was derived. Part of the intention in creating our own dictionary was to resolve some of the ambiguity and imprecision in these large dictionaries. And given the method by which we combined the three dictionaries (described above), we do expect the LSD to produce somewhat different—and ultimately improved—results.

Note that not all, indeed few, of these dictionaries provide categories clearly aimed at capturing valence or positive-negative sentiment in text. We approach each dictionary as researchers interested in capturing positive and negative tone. Thus, it is important to note that we are not always using them exactly as intended. For instance, several are constructed with multiple dimensions (that were not conducive to valence codes), which have simply been omitted. And obviously our coding of the continuous measures is somewhat crude. Nevertheless, we regard our tests as a good indication of the extent to which one can achieve a valid measure of tone using a range of currently available content-analytic dictionaries.

The measure of tone outlined below captures, simply, the degree of positive or negative coverage in news stories. Recall that the unit of analysis in the automated system is un-contextualized words, aggregated in this case to the document level. Since this process does not distinguish among words (other than accounting for their polarity), the measure necessarily reflects a combination of objective content about the various issues and events being reported on and subjective opinions or attitudes about the content itself. Positive coverage may result from attention to objectively positive events or policy successes; it might equally result from avid support for, or praise of, government policies. Likewise, negative coverage may result from attention to objectively tragic events or the failure of a policy; it might equally reflect criticism of government policies. Thus, tone should be considered a composite measure of (a) the relative negativity of the actual events or issues being covered and (b) the opinions and attitudes of newsmakers about those events and issues.[22] In terms of media and public opinion research, we expect that in many cases both components of tone matter; this general measure of the tone of coverage thus reflects both. In the event that one must distinguish between the two, researchers should clearly proceed with caution (and, more to the point, some additional data processing).

Simply comparing results across dictionaries is not enough, of course—we need to compare them with some other externally valid analysis, specifically human-coded content-analytic data. We do so here using a body of data coded by three trained human coders. The data include 900 articles from the *New York Times*. Four hundred fifty were randomly drawn from an existing database on all economics stories published in the *Times* from 1988–2008. The other 450 were randomly drawn from a previously topic-coded database of all front-page stories in the *Times* from 2007–2009. In this case, we drew 150 randomly from within each of three topic categories: environment, foreign affairs, and crime. The selection of these topics provides, we believe, a good basis for an initial test of the LSD. We have chosen a range of domestic and international policy issues; we also intentionally use issues that have seen a particularly large amount of attention in the political communication literature.

Coders were directed to read each article and then assign to each article a tone of positive, negative, or neutral. The tone is intended to reflect the overall sentiment of the article—not the tone for a particular individual or paragraph or the coders' own feelings

about the news content. Directing the coders in this general way is critical to our endeavor, since we need to have coders produce coding that is consistent with what we then ask of our dictionary.

Assigning identical codes is clearly important when we are assigning topics, or frames, but may not be as feasible where tone is concerned. As noted above, for some computational linguists, small differences across human coders are regarded as capturing real variation, or ambiguity, given the natural and structural ambiguity in categories of sentiment (Andreevskaia & Bergler, 2006; Subasic & Huettner, 2001). Following this approach, codes from the three human coders are arranged here into a 5-point scale: negative, where all three coders selected negative; mildly negative, where two coders selected negative; neutral, where two or more coders selected neutral; mildly positive, where two coders selected positive; and positive, where all three coders selected positive.[23]

The resulting distribution of stories across tone categories and topics is shown in Table 1. Overall, more than half of the stories were coded into one of the negative categories; roughly 25% of stories were coded as positive, and roughly 20% were coded as neutral. The distribution changes as we move from issue to issue, of course. Environmental coverage was comparatively positive in our database; crime and foreign affairs were somewhat more negative. This is apparent not just in the distribution of human codes, but in the automated tone measure shown in the final column of Table 1. "Net tone," our core measure of automated tone, is the proportion of positive words minus the proportion of negative words in an article, that is: (# positive words/all words) − (# negative words/all words).[24] So a score of −2.4 for crime means that, on average, in crime stories there is a 2.4-percentage-point gap between the number of negative words and the number of positive words.

How do LSD net tone scores differ across manual tone categories? This is the central test of the success of the LSD dictionary; it appears in Figure 1. The figure shows the average net tone score for each of the five codes resulting from the manual coding. Results are as we would hope—in the aggregate, more negative stories receive, in short, more negative scores. The difference in mean net tone is statistically significant, even using the more stringent two-tailed test, across all categories except one. The LSD does not distinguish especially well here between somewhat negative and neutral. There is a difference in the right direction, to be sure, but it narrowly misses statistical significance.

The issue of whether the dictionary works equally well across all four topics is the focus of Figure 2. Results are somewhat noisier, in part due to much-reduced sample sizes. And the varying tone of coverage across issues is in evidence here. The range of tone in economic articles is rather narrower than the others, and crime and foreign affairs coverage lean more toward the negative than do economic and environmental coverage (at least during the time period investigated here). Even so, all topics show the correct basic trends. The difference between neutral and somewhat negative is not as wide as we would like for economic stories or the environment, and there is a drop in tone from somewhat positive to positive in foreign affairs articles that we would not expect. But overall, we regard these results as promising.

Table 2 and Figure 3 provide some basic information comparing results using the LSD with those using other dictionaries. Table 2 makes readily apparent the fact that these dictionaries do indeed capture different things. The table shows bivariate correlations between the net tone measures that result from using one dictionary versus another. The first column is the most important for our purposes, as it shows the bivariate correlations between the LSD and all others. As expected, the correlations are relatively high between the LSD and the three dictionaries on which it is based (GI, Roget's, and RID). Roget's is slightly

Table 1
Data set descriptives

	Manual tone						
	Negative (%)	Somewhat negative (%)	Neutral (%)	Somewhat positive (%)	Positive (%)	*n*	Mean tone
All	18.45	37.23	19.12	19.57	5.62	900	−0.284
Economy	19.37	35.36	18.02	21.4	5.86	450	0.318
Crime	18.67	40.67	28	10	2.67	150	−2.414
Environment	11.03	31.03	16.55	31.72	9.66	150	0.589
Foreign	22.67	45.33	16	12	4	150	−0.836

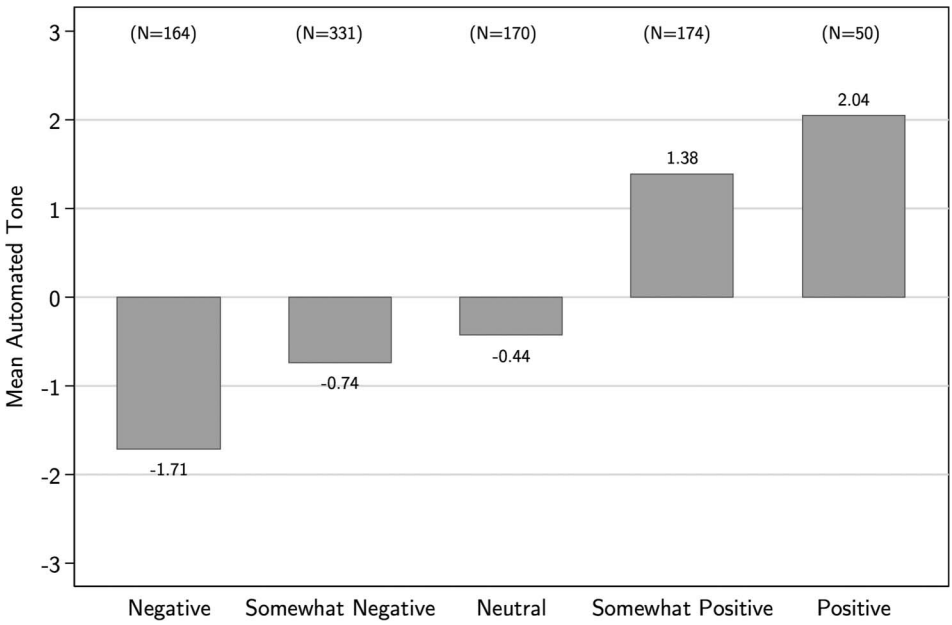

Figure 1. Automated versus manual tone.

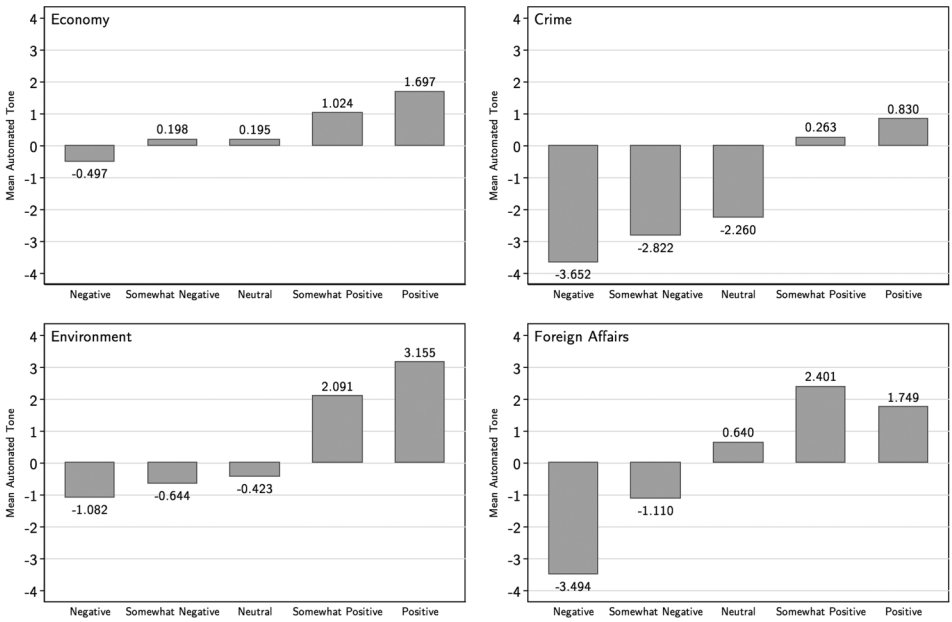

Figure 2. Comparing results across topics.

lower, and this is likely due to its breadth—many of the more arcane entries were simply not applicable for our purposes. The high correlation with LIWC makes good sense. Like the LSD, LIWC was constructed using a methodology based on definitive codes. It is one of the few to contain large positive and negative valence categories; it is also one of the

Table 2
Pairwise correlations, automated dictionaries

	LSD	GI	ROG	RID	ANEW	DAL	LIWC	PMI	TAS/C
GI	0.672								
ROG	0.471	0.469							
RID	0.669	0.480	0.350						
ANEW	0.500	0.464	0.236	0.367					
DAL	0.519	0.481	0.285	0.385	0.482				
LIWC	0.753	0.598	0.428	0.663	0.488	0.490			
PMI	0.228	0.172	0.093	0.128	0.115	0.201	0.159		
TAS/C	0.663	0.601	0.455	0.513	0.438	0.432	0.635	0.178	
WNA	0.230	0.220	0.102	0.068	0.076	0.155	0.224	0.176	0.178

Note. $N = 900$. All correlations are significant at $p < .001$.

only dictionaries making liberal use of truncation. (In our own analyses, we have found that truncation has a huge impact on performance, due to the large boost in coverage.)

Overall, bivariate correlations between results from many of these dictionaries are not especially high. This is not particularly surprising, given that the dictionaries were built for very different purposes, even as they sought to capture similar concepts. Whether they serve our purpose best is the focus of Figure 3.

Figure 3 shows directly comparable tests of external validity—replications of Figure 2, but using each of nine other content-analysis dictionaries. The scales on the y-axes vary widely, since the various dictionaries have different numbers of words in them. That the RID produces only negative values is a function of that dictionary being heavily weighted toward negative words; the opposite is true for TAS/C, which leans toward the positive. The raw values are not of primary interest here, however. What matters is whether the dictionaries produce net tone codes that systematically increase alongside results from manual coding.

In some cases, they certainly do. There is a clear and nearly monotonic increase in net tone for the GI, LIWC, and TAS/C dictionaries, and the other dictionaries perform sufficiently as well, though with somewhat rougher results. This is true in spite of the fact that the dictionaries are in some cases only barely correlated with each other and contain vastly different word lists. This highlights two facts. First, valence, as captured by manual coders, likely depends on a variety of factors, only some of which are adequately captured by any one of these dictionaries. So, two dictionaries with rather different word lists can produce aggregate code statistics that match, at least in part, human-coded results. Second, automated valence codes can be relatively successful in the aggregate even as they are noisy at the individual level. The relative success of each of these dictionaries is in part due to a large sample of articles; were we to look at any one single article, one dictionary might produce quite a different estimate of tone than another. Overall, however, the language in each of these dictionaries captures something to do with valence, and as a consequence we can find relatively sensible results in the aggregate.

How can we better judge the relative success of each of these dictionaries? Table 3 presents what we regard as the critical test. The table presents basic ANOVA results in which the variance of each net tone measure is analyzed as a function of the five-category manual results. A first version assumes a linear effect—results are essentially the R^2 values

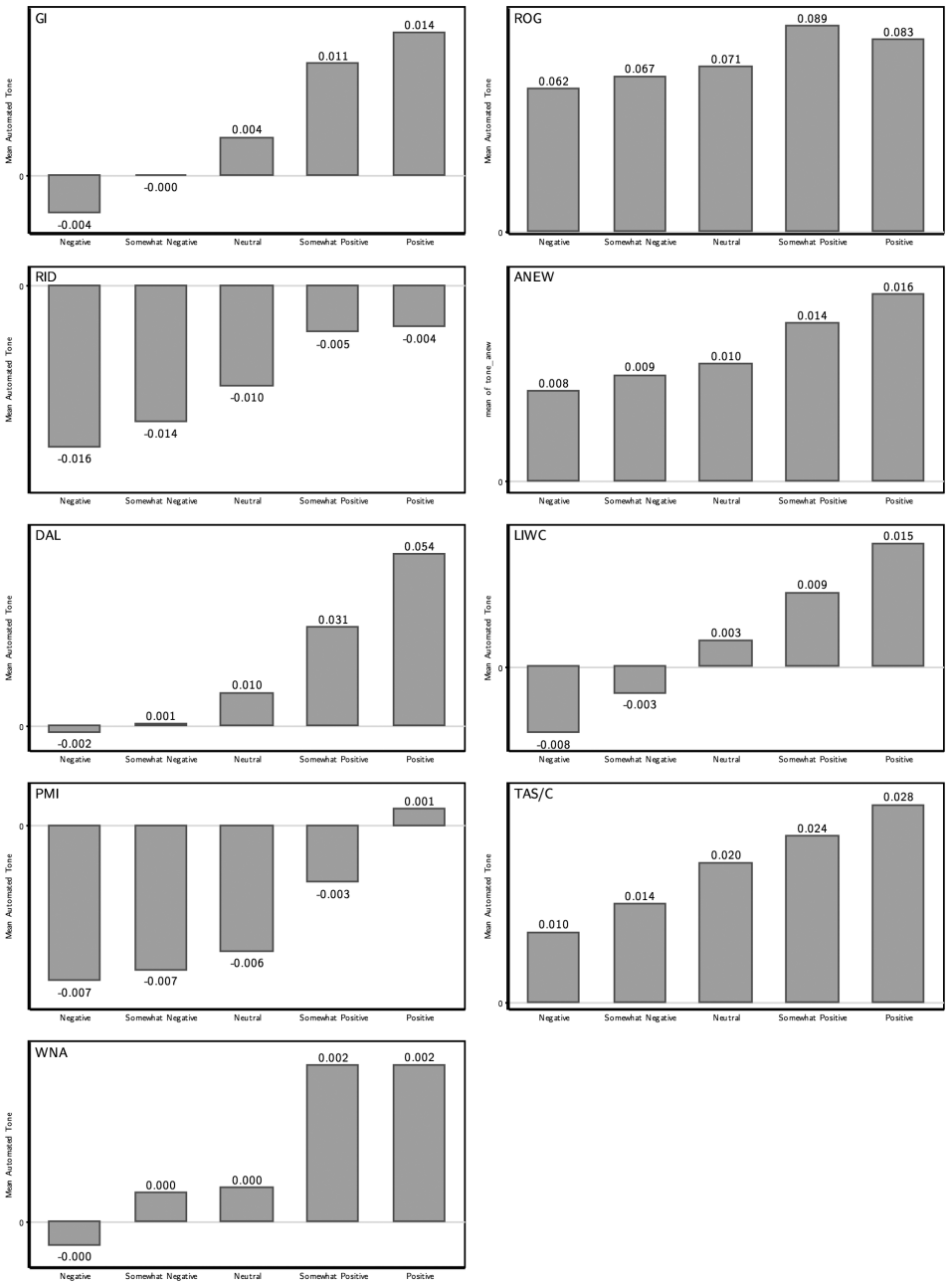

Figure 3. Comparing results across dictionaries.

from an OLS model regressing net tone on the five-category scale. A second version relaxes the assumption of linearity, and results in that case are essentially the R^2 values from a model regressing net tone on a set of four dummy variables (and one residual category) from the manual coding. Results change very little from one column to the next. In each case, the LSD matches human codes better than the other dictionaries; more precisely,

Table 3
Automated dictionaries and manual coding

	Percentage variance explained	
	Linear	Nonlinear
LSD	14.3	15.6
GI	7.0	7.2
ROG	5.3	6.2
RID	10.2	10.6
ANEW	2.3	2.5
DAL	6.4	7.3
LIWC	12.7	12.7
PMI	3.6	4.2
TAS/C	7.7	7.7
WNA	1.3	1.4

Note. Cells contain the percentage of variance in the various automated measures that is accounted for by the manual 5-point coding.

human codes account for a greater proportion of the variance in net tone as estimated using the LSD than in net tone estimated by any other means.

This is not to say that there isn't room for improvement. Results in Table 3 suggest that 15.6%[25] of the variance in net tone as determined using the LSD can be accounted for by human codes. There is a good deal of variance that is not clearly accounted for, and there surely are a good number of errors at the level of individual articles. This is actually quite a difficult thing to gauge, since the interval-level net tone measure does not have obvious cut-points at which we can easily distinguish between negative, neutral, and positive. That said, a preliminary test is illustrative. If we allow zero to be the neutral point for the LSD net tone score,[26] and then allow all net tone values that are significantly different from zero (based on our sample) to fall into either the positive or negative categories, we can compare the resulting three-way categorization of articles, as determined by the LSD, with a similar distribution using human codes. Doing so suggests that of the 224 stories categorized as positive by at least two of the three human coders, LSD results assign 74% to the positive category and just 12% to the negative category. Of the 495 articles that are categorized as negative by at least two coders, LSD results assign 53% to the negative category and 32% to the positive category.[27] Thus, it seems that the LSD performs better in the attribution of positive tone than in the attribution of negative tone. Only further testing can reveal exactly why this is the case.

Media Tone and Vote Intentions

What are the uses of the LSD? There are, we believe, many. The LSD can be used to capture the tone as well as, more generally, the use of affective language in a wide range of contexts. We explore one such context here: news about political parties and candidates in election campaigns.

Past work suggests that the tone of news content is strongly related to variations in vote intentions during election campaigns. This could be because media drive vote intentions;

it could also be that media simply reflect the tone and content of public debate at the time. Most likely, media do a little of both. In any case, the relationship between media and vote intentions tends to be rather strong.

Recent work using manually coded data points to a strong association between vote intentions and lagged media content in the 2004 and 2006 Canadian federal elections (Soroka et al., 2009). We extend this work here, comparing results from human-coded data for the 2006 election with results based on automated data using the LSD. Doing so allows us to examine not just the convergent validity of the LSD (above), but the predictive validity as well.

Models in the 2009 article predict vote shares for both the Conservative and Liberal parties in the 2006 Canadian federal election campaign, using a combination of 4-, 5-, and 6-day lags of media content (lagged in this way to allow for predictions 3 days ahead), alongside a 4-day lag of vote intentions and a set of dummy variables to capture house effects.[28] The original data were based on manually coded newspaper content, drawn directly from hard copies of newspapers during the campaign. To compare these results with automated data, we created a new database of campaign-related stories drawn from full-text indices in Nexis for the five English-language newspapers used in the original article.[29] The samples will not be identical, of course. To facilitate comparison, we matched articles by date, newspaper, and title, capturing 1,590—roughly half—of the original human-coded stories to be preprocessed and coded with the LSD. We thus rely on that subsample here to replicate the original model, and then compare results to those using the LSD.[30]

All 1,590 stories were preprocessed as above and coded for tone using the LSD. In order to attribute tone to one party/leader or the other, we look for the co-occurrence of party/leader names and positive or negative keywords in the same sentence. One consequence of this proximity-based search is that dictionary terms that occur in sentences that mention both Liberals and Conservatives are attributed to both parties. Using "net tone" for the automated measure ensures that these common words cancel each other out, which is desirable, since we do not know to whom they should be attributed. Consequently, the measure below captures the relative tone of coverage toward each actor over the campaign period.[31]

Our automated measure of leader or party tone is calculated as follows: # positive words – # negative words, using only those words that co-occur in sentences that mention the party or leader's name.[32] The measure takes on positive values when a party/leader mention co-occurs with more positive than negative words, and negative values when a party/leader mention co-occurs with more negative than positive words. Note that more party/leader mentions in a given article increase the number of words analyzed, and thus the potential value of this measure[33]; in these data, the article-level measure ranges from about −4 to 6.[34]

Results are shown in Table 4. The table includes coefficients for the media tone variables; the R^2 and adjusted R^2 values for model fit, and to assess predictive accuracy, the mean average error (MAE) of the estimate, which captures the average gap between the prediction and the actual vote intentions.[35] Control variables, including lagged vote intentions, and dummy variables capturing house effects are included in the appendix.

Model 1 is the baseline model, and includes no media variables; Model 2 includes the original manually coded media variables for each leader and party tone; and Model 3 includes the same set of media variables, though produced using the LSD. For all media variables, the table shows the summed coefficients (and standard errors) for the 4-, 5-, and 6-day lags, leaders and parties combined.

Table 4
Media content and vote intentions, 2006 Canadian election

	Conservatives			Liberals		
	1	2	3	1	2	3
Media tone coefficients						
$\sum CPC_{t-(4,5,6)}$		21.080**	−.020		−8.858	−1.161
		(7.624)	(.921)		(5.540)	(.971)
$\sum LPC_{t-(4,5,6)}$		−8.485	−1.177*		28.188**	2.047**
		(9.670)	(.636)		(6.508)	(−1.161)
Variance explained						
(R^2)	.725	.868	.840	.745	.944	.886
Accuracy						
(MAE)	1.548	1.107	1.087	1.654	.742	1.085
	(1.097)	(.709)	(.954)	(1.166)	(.587)	(.812)

Note. $N = 47$ for all models. Media tone coefficients cells contain OLS coefficients with standard errors in parentheses; MAE cells contain mean average errors with standard deviations in parentheses.
*$p < .10$; **$p < .05$.

Models 2 and 3 show the expected relationship between lagged media content and current vote intentions.[36] The variance in the automated measure is very different from that in the manual measure, so we cannot easily compare the magnitude of coefficients. We can compare the significance of coefficients and the predictive capacity of the models, however. In the Conservative party (CPC) Model 2, using manual tone, Conservative tone is positively related to vote shares, while Liberal tone is negatively signed but statistically insignificant. In Model 3, now using automated tone, Conservative tone has no significant effect, but Liberal tone is both negative and significant. The predictive capacity of the models is very similar, however. The R^2 values for the models are .868 and .840; the MAEs are 1.107 and 1.087. Clearly, the automated measure is as valuable a predictor here as the manual measure.

The Liberal models are roughly similar. Liberal tone is positive and significant in both the manual and automated models; Conservative tone is consistently negative and insignificant. In terms of model performance, the manual measure is somewhat stronger, leading to a somewhat higher R^2 value and a lower MAE. Generally speaking, there is consistency in the significance of the media variables and only a slight advantage to using the manual measure here. Overall, we find these results very encouraging.

Figure 4 confirms the close relationship between results using the manual and automated measures of party and leader tone. The panels show poll results alongside the predictions made by Models 2 and 3 in Table 4.[37] Clearly, the automated measure is capturing much of what the manual measure captures—initial evidence of the predictive validity of the LSD, at least in the context of the 2006 electoral campaign in Canada.

Conclusions

We view this study as a critical starting point for a revived discussion about dictionary construction and validation in sentiment analysis for political communication. As digital

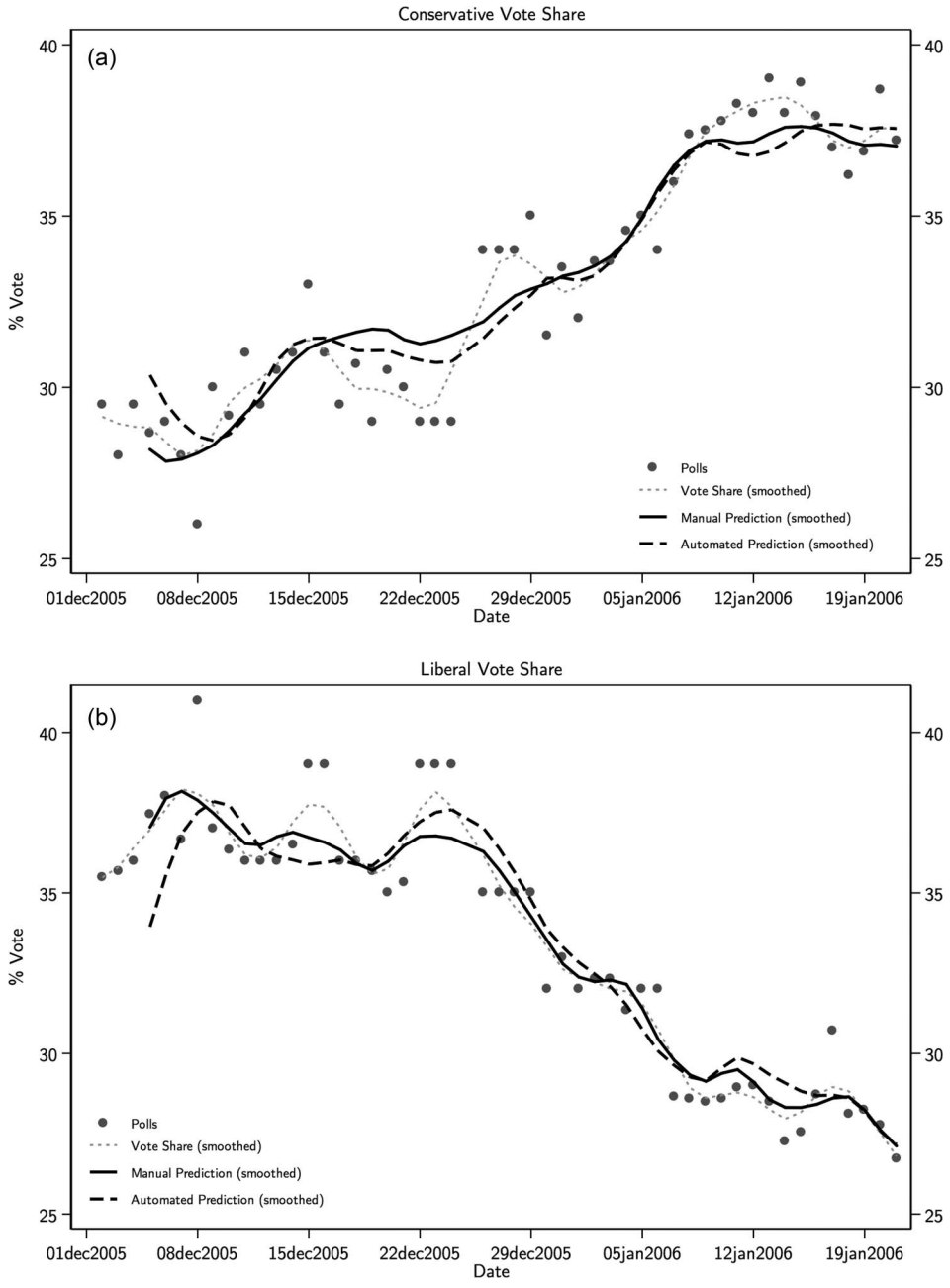

Figure 4. Media tone and vote shares in the 2006 Canadian election.

information proliferates and attention to the role of affect and emotion in politics increases, it is incumbent on us to both develop and critically evaluate relevant measurement instruments. Given the relative ease of implementing dictionary-based automation (once a dictionary is constructed), it is inevitable that these tools will become more widely available and broadly used in the near future. At a minimum, we hope to have brought attention

to the range of resources that currently exist. We hope, however, that the results above suggest the potential for the LSD to address some of the challenges of dictionary-based sentiment analysis.

Comparatively speaking, we are generally pleased with the ability of the LSD to establish the overall tone of newspaper articles. Our assessment of how well the LSD works is based on what we see as critical tests of external validity: a test of whether the dictionary produces tone codes that are in line with those produced by (expert) human coders and, moreover, a test of whether it does so more often than other available dictionaries. It appears that, in our sample, it does. This is not to say that other samples would not produce somewhat different results. Our focus here has been on political news stories, and we certainly allow for—indeed, welcome—the possibility that the dictionary will be more or less successful as it is applied to other texts.

Importantly, preprocessing accounts for a small portion of the LSD's performance, and we regard this as promising. Such modules are incredibly labor-intensive to produce (as are the dictionaries), but they will only improve with time. And ours may only have scratched the surface. For instance, we considered only one basic set of negation patterns, but linguistically there are many. In any case, by moving beyond a bag of words we have improved our confidence that even modest amounts of contextual analysis can be quite fruitful. Still, there is much room for improvement.

The dictionary also can be refined and made more efficient. An interesting finding (which we did not note above) is that it was neither the size of the dictionary (the number of words) nor the scope of coverage (the number of "hits" in a text) that drove performance. The LSD accounted for the most variance according to the ANOVA tests, though it fell squarely in the middle on both counts. Dictionaries with too many words (i.e., Roget's) or with coverage that was too vast (i.e., DAL) suffered, arguably because they lack discriminative capacity, as did those with too few words and limited scope (i.e., WNA), which may not have had enough.

The avenues for future research are many. We have tested here very broad categories of sentiment: positive and negative. There is a multitude of subcategories covering all manner of cognitive and affective concepts, however, and we hope our efforts here raise flags for anyone interested in using these more refined categories, especially across dictionaries. Operationalized as word lists, is there a difference between hope and optimism? Fear and anger? To what extent are different dictionaries measuring concepts similarly? Future research should continue to probe the development and validation of such measures so they may be used with greater confidence in scholarly work. Another promising pursuit would be in the area of subdocument lexical analysis—relating tone to particular topics or actors. Indeed, this work has begun. Although rudimentary, the proximity-based measures of actor tone used in the vote share models above make clear that attributing tone to actors using automation is both feasible and effective. Refining this technique would not only improve automated results, it would also greatly broaden the substantive research questions that automation could potentially address. Finally, while we do not directly compare the dictionary-based approach to statistical methods, such an effort would certainly be welcome.

In sum, our goal has been to contribute to the production of reliable, valid, and comparable results in automated sentiment analysis. We welcome improvements to the dictionary, and indeed a main theme of this project has been to produce a dictionary that not only is valid, but also one that is readily available to researchers interested in capturing the sentiment of news content. As we noted above, many of the existing lexicons are not readily open to the kind of improvement that comes with repeated testing by multiple researchers,

nor do they facilitate adaptation to particular research interests or goals. The LSD is, in contrast, freely available for academic use, adjustable, adaptable, and thus readily open to improvement. The current version also, based on our results above, seems to work rather well, and the potential applications of such a dictionary are rather vast. The analysis of sentiment plays a central role in work in political communication. With greater confidence in automated content-analytic measures, we will be better equipped to understand the particular ways in which the sentiment of mass media affects public opinion and behavior.

Notes

1. On the tone of news content, see discussion in the following section. On negative advertising in particular, see, for example, a meta-analysis by Lau, Sigelman, Heldman, & Babbitt (1999).

2. Affect generally refers to any conscious or unconscious feeling as distinct from a cognitive perception, and it is a necessary component of the more complex experience of emotion, which is generally conceived of as having both affective and cognitive components (Huitt, 2003). For the purposes at hand, we use the terms sentiment and tone to refer broadly to affect or emotion. Practically speaking, automation cannot distinguish between the two, and thus the distinction is adequate. Finer distinctions within these broad categories are made explicit in the text that follows.

3. On the role of emotion in politics generally, see especially Elster (1999); Hall (2002); Marcus (2000); Marcus, Neuman, and MacKuen (2000); Neumann, Marcus, Crigler, & MacKuen (2007); and Walzer (2002). On the processing of media information, see, for example, Detenber and Reeves (1996); Lang, Dhillon, and Dong (1995); and Newhagen (1998). On the primacy of affect in the decision-making process, see, for example, Abelson (1963); Damasio (1995); Huitt (2003); LeDoux (1996); Lodge and Taber (2000); and Zajonc (1984).

4. See, for example, Bloom and Price (1975); Fair (1978); Ito, Larsen, Smith, and Cacioppo (1998); Kahneman and Tversky (1979); Quattrone and Tversky (1988); and Soroka (2006).

5. For examples in computational linguistics, see, for instance, Généreux and Evans (2006); Hatzivassiloglou and McKeown (1997); Joachims (1998); Kim and Hovy (2006); Kushal, Lawrence, and Pennock (2003); Leshed and Kaye (2006); Mishne (2005); Pang et al. (2002); and Wiebe (2000). Political scientists have taken up statistical methods to automate policy positions in party manifestos (Laver, Benoit, & Garry, 2003) and the topic of congressional speeches (Purpura & Hillard, 2006).

6. For examples of UML, see, for example, Quinn, Monroe, Colaresi, and Crespin (2006); Simon and Xenos (2004); and Turney and Littman (2002).

7. There are many other examples, including Pennebaker, Slatcher, and Chung's (2005) use of word counts to derive psychological attributes of political candidates from their tone in natural conversation. Frequency analysis has also been applied to international communications to monitor conflict and various efforts to use word counts in the analysis of international relations (e.g., Doucet & Jehn, 1997; Hogenraad, 2005; Holsti, Brody, & North, 1964; Hopmann & King, 1976;). For work using DICTION specifically, see http://www.dictionsoftware.com.

8. Several studies have used proximity-based lexical rules to attribute tone to actors or topics at the subdocument level by measuring the local co-occurrence of dictionary words and a "subject" of interest (see, e.g., Mullen & Collier, 2004; Pang et al., 2002; Tong, 2001; see also research on subjectivity analysis, e.g., Wiebe, 2000). Despite many sophisticated approaches, however, state of the art machine-learning and NLP sentiment analysis techniques cannot readily unravel the topic-specific relationship between presented evidence and speaker opinion (see, e.g., Thomas et al., 2006, p. 2).

9. GI combines Osgood's Semantic Differential Scale and the Lasswell Value Dictionary.

10. The original GI software package was programmed with a set of word sense disambiguation rules that corresponded to various senses annotated in the dictionary. However, neither the program nor the rules are maintained. Thus, most research simply collapses or weighs the carefully annotated multiple word senses, to the dismay of creator Philip Stone, who laments the tendency as "a step backwards in both theory and technique" (1986, p. 76).

11. There are other approaches to word scores as well (see, e.g., Hatzivassiloglou & McKeown, 1997; Kamps, Marx, Mokken, & de Rijke, 2004; Scharl et al., 2003; Thelen & Riloff, 2002; Turney & Littman, 2002, 2003).

12. We were unfortunately unable to include other dictionaries in the construction of the LSD due to proprietary restrictions either on their use, modification, or distribution.

13. A complete list of categories classified as positive or negative is available upon request.

14. Alternately, we could have aggregated by category across dictionaries to calculate word scores. We chose to aggregate per dictionary first to avoid biasing the tone in favor of the dictionary with the most categories.

15. The General Inquirer is a notable exception (see Note 9). Hart's DICTION program also makes modest statistical adjustments by differentially weighing homographs.

16. For more on WordStat, see http://www.provalisresearch.com/wordstat/Wordstat.html.

17. Trials were conducted using a subset of subjective words, noted in the literature to improve sentiment analysis (Wiebe, 2000). However, this version did not perform as well as the full LSD.

18. These modules are available alongside the LSD online.

19. By way of example, randomly drawn positive terms include beaming, charity, cognizant, comprehend, credible, curious, dignify, dominance, ecstatic, friend, gain, gentle, justifiably, look up to, meticulous, of note, peace, politeness, reliability, and success; randomly drawn negative terms include admonish, appall, disturbed, fight, flop, grouch, huffish, hypocritical, impurity, irritating, limp, omission, oversight, rancor, relapse, sap, serpent, untimely, worrying, and yawn.

20. Lexicoder was developed by Stuart Soroka and Lori Young, and programmed by Mark Daku. It is available at http://www.lexicoder.com

21. We would very much have liked to include DICTION in our study; however, the word list—though it can be inspected—is not exportable.

22. Notably, it is not a measure of journalistic tone, bias, or subjectivity.

23. Note that differences in human codes in our sample are a matter of degree. There is no single case of both positive and negative tone codes for a single story, for instance. Thus, we are confident that differences do reflect genuine ambiguity.

24. Proportions are used to control for the varying length of articles.

25. Note that the R^2 value for unprocessed text is 1.4 points lower than above, and about 4 points lower when inflections are not applied. Full results are available from the authors.

26. It need not be, of course, and that is a first difficulty. Another option is to let the mean net tone of all neutral articles, as determined by coders, be zero. Using that value, .45, does not significantly change the results mentioned in the text.

27. The neutral category represents a real problem in this kind of analysis, since it is not clear in the LSD codes where exactly neutral ends, and the error around the mean in this particular sample is of course a very rough proxy.

28. The lengthy lag of media content (4 days or more) is a consequence of trying to build models that can predict shifts in vote shares. The original models are described in detail in Soroka et al. (2009).

29. Stories were selected by searching for the word "election" or any one of the party leader's names in the story text in searches limited by geography (Canada). Newspapers include the *Globe and Mail*, the *Toronto Star*, the *National Post*, the *Vancouver Sun*, and the *Calgary Herald*. We exclude the two French-language newspapers here, since they cannot be coded using the LSD.

30. Using this matched subsample of articles makes for a stricter test of the LSD in comparison with human coding than would a test using the entire database. This makes sense for the purposes at hand. However, we should note that this approach greatly attenuates one of the main advantages of automated coding—namely, the ability to work with much more data than could feasibly be coded by humans. Given that the reliability of automated tone will increase with sample size, and the ease with which sample size can be increased, we are limiting our automated predictions here rather severely.

31. Note that manual tone in the original study is also a measure of net tone, accounting for the relative weight of positive versus negative coverage toward each actor during the campaign.

32. Note that we include variants of party names such as "Grits" or "Tories" for Liberals and Conservatives, for instance.

33. We could also calculate net tone as a proportion of the total number of words analyzed, to account for differences in the volume of coverage of various actors. We see some advantage to the raw measure we use here, however, since it captures, in part, the consequences of a large versus small amount of negative/positive coverage. In any case, results are not very different when we use the percentage-point measure.

34. This automated measure of tone differs from the manual measure in at least one way. Recall that automated tone is a composite measure of the relative negativity of the events or issues being covered and the opinions and attitudes of newsmakers. In the original study, expert coders were trained to measure the latter. We should accordingly expect to see some differences in their relationship to vote shares. However, as the results demonstrate, any such differences turn out to be minor.

35. On the value of the MAE as a goodness of fit measure in prediction and forecasting, see Krueger and Lewis-Beck (2005).

36. And note that results from these models—relying on just 1,590 of the original manually coded articles—are not very different from the original results in Soroka et al. (2009).

37. Predictions are smoothed using lowess smoothing with a bandwidth of .2.

References

Abelson, R. P. (1963). Computer simulation of "hot" cognition. In S. Tomkins & S. Messick (Eds.), *Computer simulation of personality* (pp. 277–298). New York, NY: Wiley.

Andreevskaia, A., & Bergler, S. (2006). *Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses*. Paper presented at the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy.

Athanaselis, T., Bakamidis, S. Dologlou, I., Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). ASR for emotional speech: Clarifying the issues and enhancing performance. *Neural Networks*, *18*, 437–444.

Bloom, H. S., & Price, H. (1975). Voter response to short-run economic conditions: The asymmetric effect of prosperity and recession. *American Political Science Review*, *69*, 1240–1254.

Bolasco, S., & Ratta-Rinaldi, F. (2004). *Experiments on semantic categorisation of texts: Analysis of positive and negative dimensions*. Paper presented at the 7th International Conference on the Statistical Analysis of Textual Data, Louvain-la-Neuve, Belgium.

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings*. Gainesville: Center for Research in Psychophysiology, University of Florida.

Cho, J., Boyle, M. P., Keum, H., Shevy, M. D., McLeod, D. M., Shan, D. V., & Pan, Z. (2003). Media, terrorism, and emotionality: Emotional differences in media content and public reactions to the September 11th terrorist attacks. *Journal of Broadcasting & Electronic Media*, *47*, 309–327.

Damasio, A. (1995). *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Avon Books.

Detenber, B. H., & Reeves, B. (1996). A bio-informational theory of emotion: Motion and image size effects on viewers. *Journal of Communication*, *46*, 66–84.

Doucet, L., & Jehn, K. (1997). Analyzing harsh words in a sensitive setting: American expatriates in communist China. *Journal of Organizational Behaviour*, *18*, 559–582.

Edelman, M. (1985). Political language and political reality. *Political Science and Politics*, *18*, 10–19.

Elster, J. (1999). *Alchemies of the mind*. Cambridge, England: Cambridge University Press.

Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication*, *27*, 121–140.

Fair, R. C. (1978). The effect of economic events on votes for president. *Review of Economics and Statistics*, *60*, 159–173.

Farnsworth, S. J., & Lichter, S. R. (2010). *The nightly news nightmare: Media coverage of U.S. presidential elections, 1988–2008*. Lanham, MD: Rowman & Littlefield.

Généreux, M., & Evans, R. (2006). *Towards a validated model for affective classification of texts*. Paper presented at the Workshop of Sentiment and Subjectivity in Text, Association for Computational Linguistics, Sydney, Australia.

Gentzkow, M., & Shapiro, J. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, *78*, 35–71.

Grefenstette, G., Qu, Y., Evans, D., & Shanahan, J. (2004). Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In Y. Qu, J. Shanahan, & J. Wiebe (Eds.), *Exploring attitude and affect in text: Theories and applications* (pp. 93–107). Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Hall, C. (2002). Passions and constraint: The marginalization of passion in liberal political theory. *Philosophy and Social Criticism*, *28*, 727–748.

Hart, R. P. (1984). *Verbal style and the presidency: A computer-based analysis*. New York, NY: Academic Press.

Hart, R. P. (2001). Redeveloping diction: Theoretical considerations. In M. West (Ed.), *Theory, method, and practice in computer content analysis* (pp. 43–60). Westport, CT: Ablex.

Hart, R. P. (2000a). *DICTION 5.0: The text analysis program*. Thousand Oaks, CA: Sage-Scolari.

Hart, R. P. (2000b). *Political keywords: Using language that uses us*. New York, NY: Oxford University Press.

Hatzivassiloglou, V., & McKeown, K. (1997). *Predicting the semantic orientation of adjectives*. Paper presented at the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain.

Hogenraad, R. (2005). What the words of war can tell us about the risk of war. *Peace and Conflict: Journal of Peace Psychology*, *11*, 137–151.

Hogenraad, R., McKenzie, D., & Péladeau, N. (2003). Force and influence in content analysis: The production of new social knowledge. *Quality & Quantity*, *37*, 221–238.

Holsti, O. R., Brody, R. A., & North, R. C. (1964). Measuring affect and action in international reaction models: Empirical materials from the 1962 Cuban Crisis. *Journal of Peace Research, 1*(3/4), 170–190.

Hopmann, P. T., & King, T. (1976). Interactions and perceptions in the test ban negotiations. *International Studies Quarterly*, *20*, 105–142.

Huitt, W. (2003). *The affective system: Educational psychology interactive*. Valdosta, GA: Valdosta State University.

Ito, T. A., Larsen, J., Smith, K., & Cacioppo, J. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, *75*, 887–900.

Iyengar, S. (1996). Framing responsibility for political issues. *Annals of the American Academy of Political and Social Science*, *546*, 59–70.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European conference on machine learning* (pp. 137–142).

Johnson-Cartée, K. S. (2005). *News narrative and news framing: Constructing political reality*. Lanham, MD: Rowman & Littlefield.

Kahneman, D., & Teversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

Kamps, J., Marx, M., Mokken, R., & de Rijke, M. (2004). *Using WordNet to measure semantic orientation of adjectives*. Paris, France: European Language Resources Association.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, *22*, 110–125.

Kim, S., & Hovy, E. (2006). *Extracting opinions, opinion holders, and topics expressed in online news media text*. Paper presented at the Workshop on Sentiment and Subjectivity in Text, Sydney, Australia.

Krueger, J. S., & Lewis-Beck, M. (2005). *The place of prediction in politics*. Paper presented at the annual meeting of the American Political Science Association, Washington, DC.

Kushal, D., Lawrence, S., & Pennock, D. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. New York, NY: Association for Computing Machinery.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lang, A., Dhillon, K., & Dong, Q. (1995). The effects of emotional arousal and valence on television viewers' cognitive capacity and memory. *Journal of Broadcasting & Electronic Media*, *39*, 313–327.

Lau, R. R., Sigelman, L., Heldman, C., & Babbitt, P. (1999). The effects of negative political advertisements: A meta-analytic assessment. *American Political Science Review*, *93*, 851–875.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*, 311–331.

LeDoux, J. E. (1996). *The emotional brain*. New York, NY: Simon & Schuster.

Leshed, G., & Kaye, J. (2006). *Understanding how bloggers feel: Recognizing affect in blog posts*. New York, NY: Association for Computing Machinery.

Lodge, M., & Taber, C. (2000). Three steps toward a theory of motivated political reasoning. In A. Lupia, M. McCubbins, & S. Popkin (Eds.), *Elements of reason: Cognition, choice, and the bounds of rationality* (pp. 183–213). London, England: Cambridge University Press.

Lowry, D. T. (2008). Network TV news framing of good vs. bad economic news under Democrat and Republican presidents: A lexical analysis of political bias. *Journalism & Mass Communication Quarterly*, *85*, 483–498.

Marcus, G. E. (2002). *The sentimental citizen: Emotion in democratic politics*. University Park: Pennsylvania State University Press.

Marcus, G. E., Neuman, W., & MacKuen, M. (2000). *Affective intelligence and political judgment*. Chicago, IL: University of Chicago Press.

Martindale, C. (1975). *Romantic progression: The psychology of literary history*. Washington, DC: Hemisphere.

Martindale, C. (1990). *The clockwork muse: The predictability of artistic change*. New York, NY: Basic Books.

McComas, K., & Shanahan, J. (1999). Telling stories about global climate change: Measuring the impact of narratives on issue cycles. *Communication Research*, *26*, 30–57.

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*, *64*, 1306–1315.

Mergenthaler, E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research*, *18*, 109–126.

Mishne, G. (2005). *Experiments with mood classification in blog posts*. Paper presented at Style 2005, Stylistic Analysis of Text for Information Access, Salvador, Brazil.

Mullen, A., & Collier, N. (2004). *Incorporating topic information into sentiment analysis models*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.

Murphy, C., Bowler, S., Burgess, C., & Johnson, M. (2006). *The rhetorical semantics of state ballot initiative arguments in California, 1980–2004*. Paper presented at the American Political Science Association conference, Philadelphia, PA.

Nadeau, R., Niemi, R., Fan, D., & Amato, T. (1999). Elite economic forecasts, economic news, mass economic judgments, and presidential approval. *Journal of Politics*, *61*, 109–135.

Neumann, R., Marcus, G., Crigler, A., & MacKuen, M. (2007). *The affect effect: The dynamics of emotion in political thinking and behavior*. Chicago, IL: University of Chicago Press.

Newhagen, J. E. (1998). TV images that induce anger, fear, and disgust: Effects on approach-avoidance responses and memory. *Journal of Broadcasting & Electronic Media*, *42*, 265–276.

Nussbaum, M. (2004). *Hiding from humanity: Shame, disgust and the law*. Princeton, NJ: Princeton University Press.

Ottati, V. C., Steenbergen, R., & Riggle, E. (1992). The cognitive and affective components of political attitudes: Measuring the determinants of candidate evaluations. *Political Behavior*, *14*, 423–442.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA.

Pennebaker, J. W., Francis, M., & Booth, R. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Erlbaum.

Pennebaker, J. W., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577.

Pennebaker, J. W., Slatcher, R. B., & Chung, C. K. (2005). Linguistic markers of psychological state through media interviews: John Kerry and John Edwards in 2004, Al Gore in 2000. *Analyses of Social Issues and Public Policy*, *5*, 197–204.

Purpura, S., & Hillard, D. (2006). *Automated classification of congressional legislation*. Paper presented at the 7th Annual International Conference on Digital Government Research, San Diego, CA.

Quattrone, G., & Tversky, A. (1988). Contrasting rational and psychological analyses of political choice. *American Political Science Review*, *82*, 719–736.

Quinn, K., Monroe, B., Colaresi, M., & Crespin, M. (2006). *An automated method to topic-coding legislative speech over time with application to the 105th–109th U.S. Senate*. Paper presented at the American Political Science Association conference, Philadelphia, PA.

Roget, P. M. (1911). *Roget's thesaurus of English words and phrases* (supplemented electronic version). Salt Lake City, UT: Project Gutenberg Library Archive Foundation.

Scharl, A., Pollach, I., & Bauer, C. (2003). Determining the semantic orientation of Web-based corpora. *Lecture Notes in Computer Science*, *2690*, 840–849.

Shenhav, S. R. (2006). Political narratives and political reality. *International Political Science Review*, *27*, 245–262.

Simon, A., & Xenos, M. (2004). Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis. *Political Analysis, 12*, 63–75.

Soroka, S. N. (2006). Good news and bad news: Asymmetric responses to economic information. *Journal of Politics*, *68*, 372–385.

Soroka, S., Bodet, M., Young, L., & Andrew, B. (2009). Campaign news and vote intentions. *Journal of Elections, Public Opinion and Parties*, *19*, 359–376.

Stone, P. J. (1986). Review of Hart, R. P. 1984 *Verbal style and the presidency*: *A computer-based analysis*. New York: Academic Press. *Contemporary Sociology*, *15*(1), 75–77.

Stone, P. J., Bales, R., Namenwirth, J., & Ogilvie, D. (1962). The General Inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*, 484–494.

Stone, P. J., Dumphy, D. C., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Strapparava, C., & Valitutti, A. (2004). *WordNet-Affect: An affective extension of WordNet*. Paper presented at the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy typing. *IEEE Transactions on Fuzzy Systems*, *9*, 483–496.

Tetlock, P., Saar-Tsechansky, M., & Macskassy, S. (2007). *More than words: Quantifying language to measure firms' fundamentals.* Retrieved from http://ssrn.com/abstract=923911

Thelen, M., & Riloff, E. (2002). *A bootstrapping method for learning semantic lexicons using extraction pattern contexts*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA.

Thomas, M., Pang, B., & Lee, L. (2006). *Get out the vote: Determining support or opposition from congressional floor-debate transcripts*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.

Tong, R. (2001). *Detecting and tracking opinions in online discussions*. Paper presented at the Workshop on Operational Text Classification, New Orleans, LA.

Turney, P., & Littman, M. L. (2002). *Unsupervised learning of semantic orientation from a hundred-billion-word corpus*. Ottawa, Ontario, Canada: National Research Council of Canada.

Turney, P., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, *21*, 315–346.

Walzer, M. (2002). Passion and politics. *Philosophy and Social Criticism*, *28*, 617–633.

Whissell, C. (1989). The dictionary of affect in language. In R. Plutchnik & H. Kellerman (Eds.), *Emotion: Theory and research* (pp. 113–131). New York, NY: Harcourt Brace.

Wiebe, J. M. (2000). *Learning subjective adjectives from corpora*. Paper presented at the 17th National Conference on Artificial Intelligence, Austin, TX.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis*. Paper presented at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, British Columbia.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, *39*, 117–123.

Appendix: Media Content and Vote Intentions, 2006 Canadian Election, Control Variables

	Conservatives			Liberals		
	1	2	3	1	2	3
DV_{t-4}	.709*	.523*	.442*	.698*	.537*	.728*
	(.119)	(.126)	(.149)	(.123)	(.083)	(.106)
Strategic Council	−.410	.117	−.458	−1.580	−2.372*	−2.588*
	(.775)	(.752)	(1.205)	(.835)	(.544)	(.771)
SES	.393	1.334	.729	.633	.305	.660
	(.979)	(1.060)	(1.110)	(1.051)	(.762)	(.905)
Decima	1.150	1.196	1.947	−3.381	−2.340*	−3.173*
	(1.417)	(1.351)	(1.637)	(1.474)	(.967)	(1.252)
Ekos	2.080	2.660*	2.832*	−2.006	−2.003*	−.404
	(1.098)	(1.025)	(1.338)	(1.188)	(.767)	(1.081)
Environics	1.080	2.463	2.120	−4.305	−5.590*	.867
	(2.307)	(2.158)	(2.728)	(2.426)	(1.508)	(2.501)
Ipsos-Reid	−.741	−.356	−1.045	.424	−.413	.652
	(.920)	(.928)	(.989)	(.981)	(.669)	(.844)
Léger	−1.992	−1.169	−1.417	.154	−.896	.303
	(1.108)	(1.069)	(1.133)	(1.183)	(.745)	(1.007)
Pollara	−.444	−1.277	.213	1.956	3.276*	4.287*
	(1.746)	(1.848)	(1.871)	(1.846)	(1.309)	(1.674)
Intercept	9.888*	12.340*	18.156*	10.554	22.092*	10.896*
	(3.750)	(3.514)	(5.284)	(4.559)	(3.384)	(3.897)

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