Extracting Systematic Social Science Meaning from Text[1]

Daniel Hopkins[2] Gary King[3]

First version: March 20, 2007
This version: July 10, 2007

[1] This paper describes material that is patent pending. Our special thanks to our indefatigable undergraduate coders, Sam Caporal, Katie Colton, Nicholas Hayes, Grace Kim, Matthew Knowles, Andrew Prokop, and Keneshia Washington. Each coded numerous blogs, dealt with the unending changes we made to our coding schemes, and made many important intellectual suggestions that greatly improved our work. Matthew Knowles also helped us track down and understand the many scholarly literatures that intersected with our work, and Steven Melendez provided invaluable computer science wizardry; both are coauthors of the open source and free computer program that implements the methods described herein (ReadMe: Software for Automated Content Analysis; see http://gking.harvard.edu/readme). We thank Ying Lu for her wisdom and advice, Stuart Shieber for introducing us to the relevant computer science literature, and http://Blogpulse.com for providing us with the URLs of over a million web logs. Thanks to Ken Benoit, Doug Bond, Matt Hindman, Dan Ho, Pranam Kolari, Mark Kantrowitz, Lillian Lee, Will Lowe, Burt Monroe, Stephen Purpura, Phil Schrodt, Stuart Shulman, and Kevin Quinn for helpful suggestions or data. Thanks also to the Library of Congress (PA#NDP03-1), the Multidisciplinary Program on Inequality and Social Policy, and the Institute for Quantitative Social Science for research support. An earlier version of this paper (under a different title) was presented at the 2006 annual meetings of the Midwest Political Science Association Conference.

[2] PhD Candidate, Harvard University (Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge MA 02138; http://www.danhopkins.org, dhopkins@iq.harvard.edu, (617) 496-8300).

[3] David Florence Professor of Government, Harvard University (Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge MA 02138; http://GKing.Harvard.Edu, King@Harvard.Edu, (617) 495-2027).

Abstract

We develop two methods of automated content analysis that give approximately unbiased estimates of quantities of theoretical interest to social scientists. With a small sample of documents hand coded into investigator-chosen categories, our methods can give accurate estimates of the proportion of text documents in each category in a larger population. Existing methods successful at maximizing the percent of documents correctly classified allow for the possibility of substantial estimation bias in the category proportions of interest. Our first approach corrects this bias for any existing classifier, with no additional assumptions. Our second method estimates the proportions without the intermediate step of individual document classification, and thereby greatly reduces the required assumptions. For both methods, we also correct statistically, apparently for the first time, for the far less-than-perfect levels of inter-coder reliability that typically characterize human attempts to classify documents, an approach that will normally outperform even population hand coding when that is feasible. We illustrate these methods by tracking the daily opinions of millions of people about candidates for the 2008 presidential nominations in online blogs, data we introduce and make available with this article, and through evaluations in available corpora from other areas, including movie reviews, university web sites, and Enron emails. We also offer easy-to-use software that implements all methods described.

1 Introduction

Efforts to extract systematic meaning from text documents date to the late 1600s, when the Church, worrying about challenges to its authority, tracked the proportion of printed texts which were non-religious (Krippendorff, 2004). Similar techniques were used by early prominent social scientists, including Waples, Berelson and Bradshaw (1940, which apparently includes the first use of the term "content analysis"), and Berelson and de Grazia (1947). Content analyses like these have spread to a vast array of fields, with automated methods now joining projects based on hand coding, and have increased at least six-fold from 1980 to 2002 (Neuendorf, 2002). The recent explosive increase in web pages, blogs, emails, digitized books and articles, audio recordings (automatically converted to text), and electronic versions of formal government reports and legislative hearings and records (Lyman and Varian, 2003) suggests the potential for many new applications. Given the infeasibility of much larger scale human-based coding, the need for automated methods is growing fast. Indeed, large-scale projects based solely on hand coding have stopped altogether in some fields (King and Lowe, 2003, p.618).

The supervised learning methods we introduce in this paper take as data a potentially large set of text documents, of which a small, not necessarily random, subset is hand-coded into an investigator-chosen set of mutually exclusive and exhaustive categories.[1] As output, the methods give approximately unbiased and statistically consistent estimates of the proportion of all documents in each category. Accurate estimates of these *document category proportions* has not been a goal of most work in the classification literature, which has focused instead on increasing the accuracy of *individual document classification*. Unfortunately, methods tuned to maximize the percent of documents correctly classified can still produce substantial biases in the aggregate proportion of documents within each category.

Although individual document classifications are of great interest to scholars in computer science, statistics, text data mining, and computational linguistics, the vast majority of content analyses in social science seek to make inferences primarily about document category proportions (e.g. Mutz, 1998; Jones and Baumgartner, 2005). Thus, for example, the manager of a congressional office would find useful an automated method of sorting individual constituent letters by policy area so they can be routed to the most informed staffer to draft a response. In contrast, political scientists (and at the end of the day, the Member of Congress) would primarily be interested in the proportion of mail received in each category so they can track the intensity of constituency expression about each policy area. Policy makers or computer scientists may be interested in finding a needle in the haystack (such as a potential terrorist threat or the right web page to display from a search), but social scientists are more commonly interested in characterizing the haystack, or what we call in our title "systematic social science meaning". Of course, researchers other than social scientists are also interested in document category proportions. Individual-level classifications, when available, provide additional information to social scientists, since they enable one to aggregate in unanticipated ways, serve as variables in regression-type analyses, and help guide deeper qualitative inquiries into the nature of specific documents, but they rarely constitute the ultimate

[1] Although some excellent content analysis methods are able to delegate to the computer both the choice of the categorization scheme and the classification of documents into the chosen categories, our applications require methods where the social scientist chooses the questions and the data provide the answers. The former so-called "unsupervised learning methods" are versions of cluster analysis and have the great advantage of requiring fewer startup costs, since no theoretical choices about categories need be made ex ante and no hand coding is required (Quinn et al., 2006; Simon and Xeons, 2004). In contrast, the latter so-called "supervised learning methods," which require a choice of categories and a sample of hand coded documents, have the advantage of letting the social scientist, rather than the computer program, determine the most theoretically interesting questions (Laver, Benoit and Garry, 2003; Pang, Lee and Vaithyanathan, 2002; Kolari and Joshi, 2006). These approaches, and others such as dictionary-based methods (Gerner et al., 1994; King and Lowe, 2003), accomplish somewhat different tasks and so can often be productively used together, such as for discovering a relevant set of categories in part from the data.

quantities of interest in the conclusions of books and articles (as in Benoit and Laver, 2003).

Our first approach works with any existing document classifier and corrects for the biases that result when its classifications are aggregated into category proportions. No additional assumptions are required for this correction beyond those already made by the classifier. Our second approach also estimates the document category proportions but does not require the use of an existing method of classification or its assumptions, and indeed does not even classify individual documents as an intermediate step. This second approach drops all parametric modeling assumptions and allows the hand-coded sample to differ dramatically from the target population in both the language used and the document category frequencies. For both approaches, we also go a step further and develop a correction for the less-than-perfect levels of inter-coder reliability commonly seen in applications. This latter correction produces estimates that will normally be preferable even to hand coding all documents in the population, if that were feasible.

We begin by fixing ideas in Section 2 with our running example of blogs. We offer rules to follow in developing hand-coding schemes in Appendix A and explain how to produce numerical summaries of text, amenable to statistical analysis, in Section 3. We then discuss the different quantities of interest in text-bases analyses and our basic mathematical notation in Section 4. Section 5 discusses problems with existing approaches. We introduce our two methods in Section 6 along with empirical verification from several data sets in Section 7. Then we turn to a new correction for the usual lack of perfect intercoder reliability and misclassification in hand coding in Section 8 and summarize what can go wrong and how to avoid these problems in Section 9. Section 10 concludes.

2 Measuring Political Opinions in Blogs: A Running Example

Here, we introduce the study of blogs as our running example to explain our technology. However, our content analysis methodology is not specific to blogs and can be as easily applied to any set of unstructured, natural language text documents, such as speeches, open-ended survey responses, candidate web sites, congressional legislation, judicial opinions, newspaper editorials, company reports, private diaries, treaties, scholarly journal articles, or others.

Blogs (or "web logs") are periodic web postings, in the form of a daily diary, usually listed in reverse chronological order. Anyone may create and own a blog (for free), and she may post on it whatever she wishes. A minority of blogs are read widely whereas others are read by only a few close friends, but it is the opinions expressed not the readers we are after. Some blogs allow comments on the posts from others, but we focus in this paper only on the main posts by the blog author. Posts can be any length but typically are about a paragraph or two. Eight percent of Internet users in the U.S., or 12 million Americans, claim to have their own blog (Lenhart and Fox, 2006). The growth of blogs has been explosive, from essentially none in 2000 to estimates today that range from 39 to 100 million worldwide. Blogs are a remarkably democratic technology, with 6 million in China and 700,000 in Iran.[2]

These developments have led to the widespread view that "We are living through the largest expansion of expressive capability in the history of the human race" (Clay Shirky, quoted in Carr 2007). Blogs give individuals the ability to publish their views with a potentially worldwide audience for free, and at the same time give social scientists the ability to monitor the views of tens of millions more people on a daily basis than any previous data collection technology. What was once studied with a single snapshot of a few thousand brief survey responses can now be studied with the daily views of millions of people. Blogs capture what might have been in an earlier era

[2] See `http://www.blogpulse.com/` and `http://www.blogherald.com/2005/10/10/the-blog-herald-blog-count-october-2005/`

hallway conversations, soapbox speeches, musings of individuals in their private diaries, or purely private thoughts, and their ease of use encourages many more, and more detailed, expressions of individual opinions than ever before.

For our present purposes, we define our inferential target as the ongoing grand, national conversation about the American presidency, including specifically posts that are all or in part about President George W. Bush or any of the major contenders for the 2008 major party presidential nominations. Conversations like this have gone on throughout American history, but the development of this new technology means that for the first time ordinary Americans can participate without leaving their homes. We seek to measure on a daily basis how positive or negative the average sentiment is in the blogosphere about each politician on our list. Just like survey researchers, we have no special interest in the opinions of any specific individual, only the social science generalization about each politician, which might translate roughly into "the word on the street". The idea is to create an ongoing opinion poll that summarizes the views of people who join the national conversation to express an opinion.

Previous efforts to measure sentiments from the national conversation include more limited samples, such as studies of newspaper editorials or Sunday morning talk shows. We could easily include these text sources too, although many of the individuals involved, including politicians, journalists, and pundits now also have their own blogs. Measuring the national conversation in this way is far from the only way to define the population of interest, but it serves the methodological purpose in this paper of providing a running example, seems to be of considerable public interest, and may also be of interest to political scientists studying activists (Verba, Schlozman and Brady, 1995), the media (Drezner and Farrell, 2004), public opinion (Gamson, 1992), social networks (Huckfeldt and Sprague, 1995; Adamic and Glance, 2005), or elite influence (Zaller, 1992; Hindman, Tsioutsiouliklis and Johnson, 2003; Grindle, 2005).

We thus collect posts from highly political people who blog about politics all the time, as well as ordinary Americans who normally blog about gardening or their love lives, but choose to join the national conversation about the presidency for one or more posts. Bloggers' opinions get counted when they join the conversation by posting and not otherwise. We attempt to download and analyze all new English language blog posts every day.[3] Our specific goal is to compute the proportion of blogs in each of seven categories, including

Label	-2	-1	0	1	2
Category	extremely negative	negative	neutral	positive	extremely positive

as well as NA (no opinion) and NB (not a blog).

Although the first five categories are logically ordered, the set of all seven categories has no necessary ordering (which, e.g., rules out innovative approaches like wordscores, which at present requires a single dimension; see Laver, Benoit and Garry 2003). The NA category is a logical distinction that is separate from a neutral expressed opinion (category 0). Bloggers write in order to express opinions and so category 0 is not common although it and NA occur commonly if the blogger has it in mind to write primarily about something other than the politician we are studying. Category NB was included to ensure that the category list was exhaustive, which was especially important given the diverse nature of the web sites which get caught in web crawls. Appendix B

[3] We obtain our list of blogs by beginning with eight public blog directories and two other sources we obtained privately, including www.globeofblogs.com, http://truthlaidbear.com, www.nycbloggers.com, http://dir.yahoo.com/Computers_and_Internet/Internet/, www.bloghop.com/highrating.htm, http://www.blogrolling.com/top.phtml, a list of blogs provided by blogrolling.com, and 1.3 million additional blogs made available to us by Blogpulse.com. We then continuously crawl out from the links or "blogroll" on each of these blogs to identify our database of blogs.

gives an example of a portion of a blog post expressing an opinion about Senator Hillary Clinton for each of the ordered categories. Appendix A gives additional information about coding schemes.

This coding scheme represents an especially difficult test case both because of the mixed types included in our categorization scheme and since computer scientists have found that "sentiment categorization is more difficult than topic classification" (Pang, Lee and Vaithyanathan, 2002, p.83). In fact, blogs generally seem like a difficult case for automated content analysis, since the language used ranges from the Queen's English to "my crunchy gf thinks dubya hid the wmd's, :)!" and overall tends strongly toward the informal. In addition, blogs have little common internal structure, and nothing like the inverted pyramid format of newspaper columns. They can change in tone, content, style, and structure at any time, and the highly interactive nature of bloggers commenting on each other's blogs sometimes makes trends spread quickly through the blogosphere.[4]

We conclude this section with a preview of the type of empirical results we seek. To do this, we use the nonparametric method described below and apply it to blogosphere opinions about John Kerry before, during, and after the botched joke he told ("You know, education — if you make the most of it…you can do well. If you don't, you get stuck in Iraq.") Figure 1 gives a time series plot of the proportion of blog posts in each of the opinion categories over time. The sharp increase in the extremely negative (-2) category occurred immediately following Kerry's joke. Note also the concomitant drop in other categories occurred primarily from the -1 category, but even the proportion support in the positive categories dropped to some degree. Of course, this was one small incident in the run-up to the 2008 campaign, but it should give a sense of the widespread applicability of the methods we now describe.

3 Representing Text as Numbers

To analyze text statistically, we represent natural language as numerical variables following some standard procedures in the literature (Kolari and Joshi, 2006; Pang, Lee and Vaithyanathan, 2002; Purpura and Hillard, 2006). For example, our variable of interest summarizes an entire document (or blog post in our running example) with the document category into which it falls. Other variables are computed directly from the text and require taking three additional steps, each of which must work without human input, and all of which are designed to abstract the complexity of text to the essentials we need for further analysis. These steps may require some modification for other applications, but something like each should always be considered.

First, for practical reasons, we filter the set of documents. For example, we filter out non-English language blogs (using a filter by Cavnar and Trenkle, 1994), as well as spam blogs or "splogs" (with a technology we do not share publicly, for obvious reasons). For the purposes of this paper, we focus on blog posts about President George W. Bush (which we define as those which use the terms "Bush", "George W.", "Dubya", or "King George"). We repeated this procedure for 2008 presidential candidate Hillary Clinton (keeping blog posts which mention "Senator Clinton",

[4] Although individual blog posts can be hand coded, using hand coding to track opinions in the blogosphere in real time is infeasible. A random sample could be drawn, but since opinions sometimes change quickly over time, we would need to draw a different sample each day. Hand coding sufficient numbers each day or week is essentially impossible: either resource constraints would bind, or training quality would be sacrificed with larger numbers of coders. And even if possible, the time of those who would serve as coders, and resources used to pay them, can be productively redirected if using our approach.

Although it is an empirical question, using unsupervised learning methods to answer our specific questions in this application does not seem feasible. Applied to blogs, these methods could spontaneously produce our categories of interest or something similar, but the technique is free to choose a different categorization scheme instead. This is not always a problem, and could be highly valuable in the early stages of research when trying to decide on a coding scheme in the first place. But it is a potential problem when the social science question is clear, and especially in a case like this one where the most obvious features of blogs may be characteristics that are not of interest, such as the formality of the language or the topic of the post.

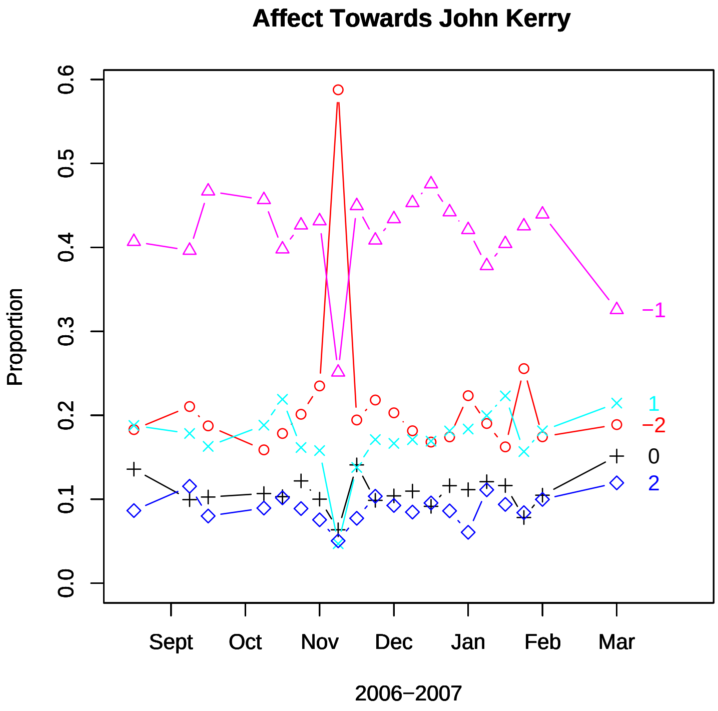

Figure 1: Blogosphere Responses to Kerry's Botched Joke. Each line gives a time series of weekly estimates of the proportion of all English language blog posts in categories ranging from −2 (extremely negative, colored red) to 2 (extremely positive, colored blue). The spike in the −2 category immediately followed Kerry's joke. Results were estimated with our nonparametric method in Section 6.2.

"Hillary", "Hitlery", and "Mrs. Clinton)," and for other candidates. Although we have downloaded millions of posts on a daily basis, we now narrow the subject for our present methodological purposes to 4,303 blog posts about President Bush collected between February 1st and February 5th, 2006, and 6,468 posts about Senator Clinton collected between August 26th and August 30th, 2006. Our method works without filtering (and even with foreign language blogs), but filters help focus the limited time of human coders on the categories of interest.

Second, we preprocess the text within each document. For example, we convert all the words to lower case and remove all punctuation marks. Another common type of preprocessing is *stemming*, which reduces some complexity by cutting the number of possible "words" we need to summarize quantitatively. For example, stemming reduces "consist", "consisted", "consistency", "consistent", "consistently", "consisting", and "consists", to their stem, which is "consist". Stemming documents strips out information, in addition to reducing complexity, but long experience in this literature is that the trade off seems well worth it (Quinn et al., 2006; Pang, Lee and Vaithyanathan, 2002; Porter, 1980).

Finally, we summarize the preprocessed text as a set of dichotomous variables, one type for the presence or absence of each word stem (or "unigram"), a second type for each word pair in a given sequence (or "bigram"), a third type for each word triplet in sequence (or "trigram"), and so on to all "n-grams". This definition is not limited to word stems that appear in dictionaries. In our application, we measure only the presence or absence of stems rather than counts (the second time the word "awful" appears in a blog post does not add anywhere near as much information as the first). For applications with larger documents, counts may be more informative and could also be included. However, even with this abstraction, the number of variables remaining is astounding — orders of magnitude larger than the number of blogs. For example, our sample of 10,771 blog posts about President Bush and Senator Clinton includes 201,676 unique unigrams, 2,392,027 unique bigrams, and 5,761,979 unique trigrams. This confirms that bloggers will not run out of new ways of saying things any time soon, but it also means we need some further simplification. The usual choice is to consider only dichotomous stemmed unigram indicator variables (the presence or absence of each of a list of word stems), which we have found to work well for the examples we have studied. We also put more emphasis on variables that discriminate by deleting stemmed unigrams appearing in fewer than 1% or greater than 99% of all documents, which results in only 3,672 variables. These procedures effectively group the infinite range of possible blog posts to "only" $2^{3,672}$ distinct types. This makes the problem feasible but still represents a huge number (larger than the number of elementary particles in the universe).

Many other types of preprocessing can also be considered, but which we found unnecessary for our particular applications. For example, one can code variables to represent the metadata, or information about the text document that is not strictly part of the text. For our application, the URL of the blog, the title, or the blogroll may convey some additional information. For example, we could have coded blogrolls by whether they cite the top 100 liberal or top 100 conservative blogs as an indication of the partisan identification of the blogger. Other possibilities include measures for whether each word stem appears near the start or end of the text, a procedure to tag each word with a part of speech, to attempt to include bigrams when unigrams will lose too much meaning, such as by replacing the two unigrams "white house" with "white_house" (Das and Chen, 2001), or to include information from the linkage structure among the set of blogs (Thomas, Pang and Lee, 2006). These additions and others provide important advantages in some applications. Overall, the conclusion of the literature is that a brute force unigram-based method, with rigorous empirical validation, will typically account for the majority of the available explanatory power.

4 Notation and Quantities of Interest

Our procedures require two sets of text documents. The first is a small *labeled set*, for which each document i ($i = 1, \ldots, n$) is hand-coded into, or somehow otherwise labeled with, one of the given categories from our categorization scheme (we discuss how large n needs to be in Section 7). We denote the Document category variable as D_i, which in general takes on the value $D_i = j$, for possible categories $j = 1, \ldots, J$.[5] (In our running example, D_i takes on the potential values $\{-2$, -1, 1, 0, 1, 2, NA, NB$\}$.) We denote the second, larger set of documents as the target *population*, in which each document ℓ (for $\ell = 1, \ldots, L$) has an *unobserved* classification D_ℓ. Sometimes the labeled set is a sample from the population and so the two overlap; more often it is a nonrandom sample from a different source than the population, such as from earlier in time.

The user need not provide any other variables, as everything else is computed directly from the documents. To define these variables for the labeled set denote S_{ik} as equal to 1 if word Stem k ($k = 1, \ldots, K$) is used at least once in document i (for $i = 1, \ldots, n$) and 0 otherwise (and similarly for the population set substituting index i with index ℓ). This makes our abstract summary of the text of document i the set of these variables, $\{S_{i1}, \ldots, S_{iK}\}$, which we summarize as the $K \times 1$ vector of word stem variables $\boldsymbol{S}_i$. We refer to $\boldsymbol{S}_i$ as a *word stem profile* since it provides a summary of all the word stems under consideration used in a document. This vector can also include other features of the text such as based on any n-gram, or variables coded from the metadata.

[5] This notation is from King and Lu (2007), who use related methods applied to unrelated substantive applications that do not involve coding text, and different mnemonic associations.

The quantity of interest in most of the supervised learning literature is the set of individual classifications for all documents in the population:

$$\{D_1, \ldots, D_L\}. \tag{1}$$

In contrast, the quantity of interest for most content analyses in social science is the aggregate proportion of all (or a subset of all) of these population documents that fall into each of the document categories:

$$P(D) = \{P(D = 1), \ldots, P(D = J)\}' \tag{2}$$

where $P(D)$ is a $J \times 1$ vector, each element of which is a proportion computed by direct tabulation:

$$P(D = j) = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}(D_\ell = j), \tag{3}$$

where $\mathbf{1}(a) = 1$ if a is true and 0 otherwise.

Document category D_i is one variable with many possible values, whereas word profile $\boldsymbol{S}_i$ constitutes a set of dichotomous variables. This means that $P(D)$ is a multinomial distribution with J possible values and $P(\boldsymbol{S})$ is a multinomial distribution with 2^K possible values, each of which is a possible word stem profile.

5 Issues with Existing Approaches

We discuss in this section the problems with two existing methods that can be used to estimate social aggregates rather than individual classifications. These include (1) *direct sampling* and (2) the *aggregation of individual document classifications* produced by supervised learning algorithms. We show how accurate estimation depends crucially on the quantity of interest, and how most methods in the literature optimize goals other than those of most interest to social scientists.

5.1 Existing Approaches

Perhaps the simplest method of directly estimating $P(D)$ is to identify a well-defined population of interest, draw a random sample from the population, hand code all the documents in the sample, and tabulate the hand-coded documents into each category. Drawing proper inferences with this method requires only basic sampling theory. It does not even require abstract numerical summaries of the text of the documents such as word stem profiles or classifications of individual documents in the population set.

The second approach to estimating $P(D)$ is standard in the supervised learning literature. The idea is to first use the labeled sample to estimate a functional relationship between document category D and word features $\boldsymbol{S}$. Typically, D serves as a multicategory dependent variable and is predicted with a set of explanatory variables $\{S_{i1}, \ldots, S_{iK}\}$, using some statistical (or machine learning, or rule-based) method. Then the coefficients of the model are estimated, and both the coefficients and the data generating process are assumed the same in the labeled sample as in the population. The coefficients are then ported to the population and used with the features measured in the population, $\boldsymbol{S}_\ell$, to predict the classification for each population document D_ℓ. Social scientists who use these approaches follow the same steps and then aggregate the individual classifications via Equation 3 to estimate their quantity of interest, $P(D)$. Many models have been chosen to perform the basic classification task, including regression, discriminant analysis, radial basis functions, multinomial logit, CART, random forests, neural networks, support vector machines, maximum entropy, and others.

5.2 Problems

Unfortunately, as Hand (2006) points out, the standard supervised learning approach to individual document classification will fail in two circumstances, both of which appear common in practice. And even if classification succeeds with high or optimal accuracy, the next subsection shows that estimating population proportions can still be biased.

When the labeled set is not a random sample from the population, both methods fail. Yet "in many, perhaps most real classification problems the data points in the [labeled] design set are not, in fact, randomly drawn from the same distribution as the data points to which the classifier will be applied. . . . It goes without saying that statements about classifier accuracy based on a false assumption about the identity of the [labeled] design set distribution and the distribution of future points may well be inaccurate" (Hand, 2006, p.2). Deviations from randomness may occur due to what Hand calls "population drift," which occurs when the labeled set is collected at one point and meant to apply to a population collected over time (as in our blogs example), or for many other reasons.

The lack of random sampling would seem to be an even more common characteristic of real social science applications, which have many aggregate quantities of interest. A single hand-coded data set is typically insufficient to estimate all the quantities. That is, a study that asks only a single question is rare. Almost all social science analyses also study their questions within subdivisions of their population of interest. The subdivisions may include time periods to help identify trends or, to seek out other patterns, they may include subdivisions by policy areas, speakers, countries, income groups, partisan identification, or others. If we could draw a separate random sample from each subdivision, we could estimate each separate $P(D)$ by direct sampling, but the burdens of hand coding would quickly overwhelm any researcher's coding capacity. And even in the unlikely case where we could collect a random sample for each, scholars continually develop new questions, and thus new quantities of interest, quicker than any coding team could respond.

The second failure condition is more subtle but more insidious. The data generation process assumed by the standard supervised learning approach predicts D with $\boldsymbol{S}$, modeling $P(D|\boldsymbol{S})$. However, this is not the way the world works. To take our running example, bloggers do not start writing and only afterwords figure out their affect toward the president: they start with a view, which we abstract as a document category, and then sit down to set it out in words. That is, the right data generation process is the inverse of what is being modeled, where we should be predicting $\boldsymbol{S}$ with D, and inferring $P(\boldsymbol{S}|D)$. The consequence of using $P(D|\boldsymbol{S})$ instead (and without Bayes Theorem, which is not very helpful in this case) is the requirement of two assumptions needed to generalize from the labeled sample to the population. The first assumption is that $\boldsymbol{S}$ "spans the space of all predictors" of D (Hand, 2006, p.09), which means that once you control for your measured variables, there exists no other variable that could improve predictive power at all. In problems involving human language, including our running example, this assumption is virtually never met, since $\boldsymbol{S}$ is intentionally an abstraction of the content of the document and so by definition does not represent all existing information. As such, $\boldsymbol{S}$ does *not* span the space of all predictors. The other assumption is that the class of models chosen for $P(D|\boldsymbol{S})$ includes the "true" model. This is a more familiar assumption to social scientists, but it is of course no easier to meet. In this case, finding even the best model or a good model, much less the "true one," would be extraordinarily difficult and time consuming given the huge number of potential explanatory variables coded from text in unigrams, bigrams, etc.

As Hand (2006, p.9) writes, "Of course, it would be a brave person who could confidently assert that these two conditions held." We show how to avoid each of these impossible assumptions even without a labeled set that is a random sample from the population.

5.3 Optimizing for a Different Goal

Here we show that even optimal individual document classification that meets all the assumptions of the last section can lead to biased estimates of the document category proportions of interest to social scientists. The criteria for success in the classification literature is the percent correctly classified in a true test set. This is one reasonable criterion when the focus is on individual-level classification, but it is sometimes insufficient even if the goal is individual classification and can be seriously misleading for the general social science purpose of characterizing document populations. For example, of the 23 models estimated in the prominent analysis of Pang, Lee and Vaithyanathan (2002), the percent correctly predicted ranged from 77% to 83%. This is an excellent classification performance for the difficult problem of sentiment analysis they analyzed, but suppose that all the misclassifications were in a particular direction for one or more categories. In that situation, the statistical *bias* (the average difference between the true and estimated proportion of documents in a category) in using this method to estimate the aggregate quantities of interest could be as high as 17 to 23 percentage points. This should not matter for the authors, since their goal was classification, but it could matter for social scientists interested in category proportions.

The key problem is that, except at the extremes, there exists no necessary connection between low misclassification rates and low bias: It is easy to construct examples of learning methods that achieve a high percent of individual documents correctly predicted and large biases for estimating the aggregate document proportions, or other methods that have a low percent correctly predicted but nevertheless produce relatively unbiased estimates of the aggregate quantities. For example, flipping a coin is not a good predictor of which party will win a presidential election, but it does happen to provide an unbiased estimate of the percentage of Democratic Presidential victories since the first World War. Evidence on bias might be useful for individual classification but is essential for estimating aggregate quantities of interest. Yet, since authors in this literature are interested primarily in individual classification, they do not often report the different types of misclassification errors their methods produce or bias in estimating the aggregates. As such, the bulk of the supervised learning classification literature offers no indication of whether the methods proposed would work well for social scientists and their specific goals. Fortunately, this problem is easy to overcome, which we do in our applications.

6 Statistically Consistent Estimates of Social Aggregates

We now introduce two methods optimized for estimating document category proportions. These methods represent not merely a different focus, but also an opportunity to improve estimation accuracy substantially with far less onerous assumptions. The first method corrects aggregations of any existing classification method, whereas the second is a stand-alone procedure not requiring or providing individual document classifications.

6.1 Corrected Aggregations of Individual Classifications

Intuition Consider multinomial logit or any other method offered in the supervised learning literature to make individual-level classification decisions. Fit this model to the labeled set, use it to classify each of the unlabeled documents in the population of interest, and aggregate the classifications to obtain a raw, uncorrected estimate of the proportion of documents in each category. (This is the approach used by almost all social scientists who use classification techniques to study document category proportions.)

Next, estimate misclassification probabilities by first dividing the labeled set of documents into a training set and a test set (ignoring the unlabeled population set). Then apply the same classification method to the training set alone and make predictions for the test set, $\hat{D}_i$ (ignoring the test set's labels). Then use the test set's labels to calculate the specific misclassification probabilities between each pair of actual classifications given each true value, $P(\hat{D}_i = j|D_i = j')$. These misclassification probabilities are of little or no help in improving the individual document classifications, since they do not tell us which documents are misclassified. However, they can be used to correct the raw estimate of the document category proportions.

For example, suppose we learn, in predicting the test set proportions from the training set, that 17% of the documents our method classified as $D = 1$ really should have been classified as $D = 3$. For any one individual classification in the population, this fact is of no help. But for document category proportions, its easy to use: subtract 17% from the raw estimate of the category 1 proportion, $P(D = 1)$, and add it to category 3, $P(D = 3)$. Even if the raw estimate was badly biased, which can occur even with optimal individual document classification, the resulting corrected estimate would be unbiased and statistically consistent so long as the population misclassification errors were estimated well enough from the labeled set (a condition we discuss more below). Even if the percent corrected predicted is low, this corrected method can give unbiased estimates of the aggregate document category frequencies.

Formalization for Two Categories For the special case where D is dichotomous, the misclassification correction discussed in the previous section is well known in epidemiology — an area of science directly analogous to the social sciences, where much data are at the individual level, but the quantities of interest are usually at the population level. To see this, temporarily consider a dichotomous D, with values 1 or 2, a raw estimate of the proportion of documents in category 1 from some method of classification, $P(\hat{D} = 1)$, and the true proportion (corrected for misclassification), $P(D = 1)$.[6] Then define two forms of correct classification as "sensitivity," sens $\equiv P(\hat{D} = 1|D = 1)$ (sometimes known as "recall"), and "specificity," or spec $\equiv P(\hat{D} = 2|D = 2)$. For example, sensitivity is the proportion of documents that were predicted to be in category 1 among those actually in category 1.

Then we note that the proportion of documents estimated to be in category 1 must come from only one of two sources: documents actually in category 1 that were correctly classified and documents in actually category 2 but misclassified into category 1. We represent this accounting identity, known as the Law of Total Probability, as

$$P(\hat{D} = 1) = (\text{sens})P(D = 1) + (1 - \text{spec})P(D = 2). \tag{4}$$

Since Equation 4 is one equation with only one unknown [since $P(D = 1) = 1 - P(D = 2)$], it is easy to solve. As Levy and Kass (1970) first showed, the solution is

$$P(D = 1) = \frac{P(\hat{D} = 1) - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}. \tag{5}$$

This expression can be used in practice by estimating sensitivity and specificity in the first stage analysis (separating the labeled set into training and test sets as discussed above or more formally by cross-validation), and using the entire labeled set to predict the (unlabeled) population set to give $P(\hat{D} = 1)$. Plugging in these values in the right side of (5) gives a corrected, and statistically consistent, estimate of the true proportion of documents in category 1.

[6] The raw estimate $P(\hat{D} = 1)$ can be based on the proportion of individual documents classified into category 1. However, a better estimate for classifiers that give probabilistic classifications is to sum the estimated probability that each document is in the category for all documents. For example, if 100 documents each have a 0.52 probability of being in category 1, then all individual classifications are into this category. However, since we would only expect that 52% of documents to actually be in category 1, a better estimate is $P(\hat{D} = 1) = 0.52$.

Generalization to Any Number of Categories The applications in epidemiology for which these expressions were developed are completely different than our problems, but the methods developed there are directly relevant. This connection enables us to use for our application the generalizations developed by King and Lu (2007).[7] They show how to generalize Equation 4 to include any number of categories. This is done by substituting j for 1, and summing over all categories instead of just 2:

$$P(\hat{D} = j) = \sum_{j'=1}^{J} P(\hat{D} = j|D = j')P(D = j') \tag{6}$$

Given $P(\hat{D})$ and the misclassification probabilities, $P(\hat{D} = j|D = j')$ which generalize sensitivity and specificity to multiple categories, this expression represents a set of J equations (i.e., defined for $j = 1, \ldots, J$) that can be solved for the J elements in $P(D)$. This is aided by the fact that the equations include only $J - 1$ unknowns since elements of $P(D)$ must sum to 1.

Interpretation Section 5.3 shows that a method which meets all the assumptions required for optimal classification performance (given in Section 5.2), and which therefore maximizes the percent of individual documents correctly classified, can still give biased estimates of the document category proportions. In this section, we therefore offer statistically consistent estimates of document category proportions even without having to improve individual classification accuracy.

The method introduced here requires no assumptions beyond those already made by the individual document classifier. In particular, classifiers require that the labeled set be a random sample from the same population as the unlabeled target set of documents; the same assumption guarantees that our correction will work. More specifically, our method only requires a special case of the random selection assumption: that the misclassification probabilities (sensitivity and specificity with 2 categories or $P(\hat{D} = j|D = j')$ for all j and j' in Equation 6) estimated with data from the labeled set also hold in the unlabeled population set. This assumption may be wrong, but if it is, then the assumptions necessary for the original classifier to work are also wrong and will not necessarily even give accurate individual classifications.

The next section, which shows how to avoid individual classification altogether, works with less demanding assumptions.

6.2 Document Category Proportions Without Individual Classifications

We now offer a second approach requiring no parametric statistical modeling, individual document classification, or random sampling from the target population. It also correctly treats S as a consequence rather than cause of D. The resulting method has much less stringent assumptions necessary for accurate estimation.

[7] King and Lu's paper contributed to the field in epidemiology called "verbal autopsies." The goal of this field is to estimate the distribution of the causes of death in populations without medical death certification. This information is crucial for directing world health dollars to the problem areas. Data come from two sources. One is a sample of deaths from the population, where a relative of each deceased is asked a long (50–100 item) list of usually dichotomous questions about symptoms the deceased may have suffered prior to death ($\boldsymbol{S}_\ell$). The other source of data is deaths in a nearby hospital, where the same data collection of symptoms from relatives are collected ($\boldsymbol{S}_i$) and also where medical death certification is available (D_i). Their method produces approximately unbiased and consistent estimates, considerably better than the existing approaches which included expensive and unreliable physician reviews (where three physicians spend 20 minutes with the answers to the symptom questions from each deceased to decide on the cause of death), reliable but inaccurate expert rule-based algorithms, or model-dependent parametric statistical models.

The Method

This method only requires only one additional step beyond that in Section 6.1. Thus, instead of some statistical or machine learning method that uses $\boldsymbol{S}$ and D to estimate $P(\hat{D} = j)$, and then correcting via Equation 6, we follow (King and Lu, 2007) and recognize that we can write an analogous equation using $\boldsymbol{S}$ in place of $\hat{D}$:

$$P(\boldsymbol{S} = s) = \sum_{j=1}^{J} P(\boldsymbol{S} = s|D = j)P(D = j). \tag{7}$$

That is, any observable implication of the true D can be used in place of $\hat{D}$, and since $\hat{D}$ is a function of $\boldsymbol{S}$ — since the words chosen are by definition a function of the document category — it certainly can be used. To simplify, we rewrite Equation 7 as an equivalent matrix expression:

$$\underset{2^K \times 1}{P(\boldsymbol{S})} = \underset{2^K \times J}{P(\boldsymbol{S}|D)}\underset{J \times 1}{P(D)} \tag{8}$$

where, as indicated, $P(\boldsymbol{S})$ is the probability of each of the 2^K possible word stem profiles occurring,[8] $P(\boldsymbol{S}|D)$ is the probability of each of the 2^K possible word stem profiles occurring within the documents in category D (columns of $P(\boldsymbol{S}|D)$ corresponding to values of D), and $P(D)$ is our J-vector quantity of interest.

Elements of $P(\boldsymbol{S})$ can be estimated by direct tabulation from the target population, without parametric assumptions: one merely computes the proportion of documents observed with each pattern of word profiles. Since D is not observed in the population, we cannot estimate $P(\boldsymbol{S}|D)$ directly. Instead, we make the crucial assumption that its value in the labeled, hand-coded sample, $P^h(\boldsymbol{S}|D)$, is the same as that in the population,

$$P^h(\boldsymbol{S}|D) = P(\boldsymbol{S}|D), \tag{9}$$

and use the labeled sample to estimate this matrix (we discuss this assumption below). We avoid parametric assumptions here too, by using direct tabulation to compute the proportion of documents observed to have each specific word profile among those in each document category.

In principle, we could estimate $P(D)$ in Equation 8 assuming only the veracity of Equation 9 and the accuracy of our estimates of $P(\boldsymbol{S})$ and $P(\boldsymbol{S}|D)$, by solving Equation 8 via standard regression algebra. That is, if we think of $P(D)$ as the unknown "regression coefficients" β, $P(\boldsymbol{S}|D)$ as the "explanatory variables" matrix X, and $P(\boldsymbol{S})$ as the "dependent variable" Y, then Equation 8 becomes $Y = X\beta$ (with no error term). This happens to be a linear expression but not because of any assumption imposed on the problem that could be wrong. The result is that we can solve for $P(D)$ via the usual regression calculation: $\beta = (X'X)^{-1}X'y$. A key point is that that this calculation does *not* require classifying individual documents into categories and then aggregating; it estimates the aggregate proportions directly.

This simple approach runs into two difficulties that requires fixing in our application. First, K is typically very large and so 2^K is far larger than any computer could handle. Second is a sparseness problem since the number of observations available to tabulate for estimating $P(\boldsymbol{S})$ and $P(\boldsymbol{S}|D)$ is much smaller than the number of potential word profiles ($n << 2^K$). To avoid both of these issues, we adapt results from King and Lu (2007) and randomly choose subsets of between approximately 5 and 25 words (the number being chosen by cross-validation within the labeled set), solve for $P(D)$ in each, and average the results. Because $\boldsymbol{S}$ is treated as a consequence

[8] For example, if we ran the method with only $K = 3$ word stems, $P(\boldsymbol{S})$ would contain the probabilities of each of these ($2^3 = 8$) patterns occurring in the set of documents: 000 (i.e., none of the three words were used), 001, 010, 011, 100, 101, 110, and 111.

of D, using subsets of $\boldsymbol{S}$ introduces no new assumptions. This simple subsetting procedure turns out to be equivalent to a version of the standard approach of smoothing sparse matrices via kernel densities although, unlike the typical use of this procedure, its application here reduces bias. Standard errors and confidence intervals are computed via bootstrapping. These and other computational and statistical issues, such as constraining $P(D)$ to the simplex, are developed in King and Lu (2007).

Interpretation

A key advantage of estimating $P(D)$ directly without the intermediate step of computing the individual classifications is that the assumptions required to make it work are remarkably less restrictive. The necessary assumptions can still be wrong, and as a result our estimates can be biased, but the dramatic reduction in their restrictiveness means that under the new approach we have a fighting chance to get something close to the right answer in many applications where valid inferences were not previously likely.

As described in Section 5, to apply the direct sampling or standard supervised learning approaches, the labeled document set must be a random sample from the target population; the set of word stem profiles must span all the predictive information in the documents; and the class of parametric models chosen must include something close to the "true" data generation process. Primarily since the data generation process is $P(\boldsymbol{S}|D)$ but these models are based on $P(D|\boldsymbol{S})$, satisfying these assumptions in real data would be unlikely.

In contrast, our approach allows the distribution of documents across word stem profiles, $P(\boldsymbol{S})$, and the distribution of documents across the categories, $P(D)$, to each be completely different in the labeled set and population set of documents. So for example, if a word or pattern of words becomes more popular between the time the labeled set was hand coded and the population documents were collected — or neologisms appear on the scene, such as "wmd's", "troop surge," or "blowback" — no biases would emerge. Similarly, if documents in certain categories became more prevalent in the population than labeled set, no biases would be created. In our running example, no bias would be induced if the labeled set includes a majority of conservative Republicans who defend everything President Bush does and the target population has a super-majority of liberal Democrats who want nothing more than to end the Bush presidency. In contrast, changes in either $P(D)$ or $P(\boldsymbol{S})$ between the labeled and population sets would be sufficient to doom any of the existing classification-based approaches. For example, so long as "idiot" remains an insult, our method can make appropriate use of that information, even if the word becomes less common (a change in $P(S)$) or if there are fewer people who deserve it (a change in $P(D)$).

The key theoretical assumption of the new approach is Equation 9 — that the language used to describe a particular document category is the same in both samples. To be more specific, among all documents in a given category, the prevalence of particular word profiles in the labeled set should be the same in expectation as in the population set. For example, the language bloggers use to describe an "extremely negative" view of Hillary Clinton in the labeled set must at least be a subset of the way she is described in the target population. They do not need to literally write the same blog posts, but rather need to have the same probabilities of using similar word profiles so that $P^h(\boldsymbol{S}|D = -2) = P(\boldsymbol{S}|D = -2)$. This assumption can be violated due to population drift or for other reasons, but we can always hand code additional cases in the population set to verify that it holds sufficiently well. The number of examples of each category need not be the same either in the two document sets. And as discussed above, the proportion of examples of each document category can differ between the two document sets, although in practice since we estimate the elements of $P(\boldsymbol{S}|D)$ nonparametrically, we need sufficient examples of documents within each category in the labeled set.

Applying this methodology has the advantage not only of requiring fewer and less restrictive assumptions but also of being considerably easier to use in practice. Applying the standard supervised learning approach is difficult, even if we are optimistic about meeting its assumptions. Forget about choosing the "true" model: merely finding a "good" specification with thousands of explanatory variables to choose from can be extraordinarily time consuming. One needs to fit numerous statistical models, consider many specifications within each model type, run cross-validation tests, and check various fit statistics. Social scientists have a lot of experience with specification searches, but all the explanatory variables mean that even one run would take considerable time and many runs would need to be conducted.

The problem is further complicated by the fact that social scientists are accustomed to choosing their statistical specifications in large part on the basis of prior theoretical expectations and results from past research, whereas the overwhelming experience in the information extraction literature is that radically empirical approaches work best for a given amount of effort. For example, we might think we could carefully choose words or phrases to characterize particular document categories (e.g., "awful", "irresponsible," "impeach" etc., to describe negative views about President Bush), and indeed this approach will often work to some degree. Yet, a raw empirical search for the best specification, ignoring these theoretically chosen words, will typically turn up predictive patterns we would not have thought of ex ante. Indeed, methods based on highly detailed parsing of the grammar and sentence structure in each document can also work exceptionally well (e.g. King and Lowe, 2003), but the strong impression from the literature is that the extensive, tedious work that goes into adapting these approaches for each application are more productively put into collecting more hand-coded examples and then using an automatic specification search routine.

7 The Estimator in Practice

In this section, we show how our approach works in practice. We begin with a simple simulated example, then proceed to several real examples from different fields, and conclude with an empirical examination of how many documents one needs to hand code to make this method work.

7.1 Simulations from Blog Data

We begin with a simulated data set of 10 words and thus $2^{10} = 1,024$ possible word stem profiles. We set the elements of $P^h(D)$ to be the same across the seven categories, and then set the population document category frequencies, $P(D)$, to very different values. We assume that the same distribution of language used for each category is approximately the same in the two sets, so that Equation 9 holds. We then draw a value $\tilde{D}$ from $P^h(D)$, insert the simulation into $P^h(\boldsymbol{S}|\tilde{D})$ and then draw the simulated matrix $\tilde{\boldsymbol{S}}$ from this density. We repeat this 1,000 times to produce the labeled data set, and analogously for the population.

Figure 2 summarizes the sharp differences between the hand coded and population distributions in these data. The left graph plots $P^h(D)$ horizontally by $P(D)$ vertically, where the seven circles represent the category proportions. If the proportions were equal, they would all fall on the $45°$ line. If one used the labeled, hand-coded sample in this case via direct sampling to estimate the document category frequencies in the population, the result would not even be positively correlated with the truth.

The differences between the two distributions of word frequency profiles appear in the right graph (where for clarity the axes, but not labels, are on the log scale). Each circle in this graph represents the proportion of documents with a specific word profile. Again, if the two distributions were the same, all the circles would appear on the diagonal line, but again many of the circles fall far from this line, indicating the large differences between the two samples.

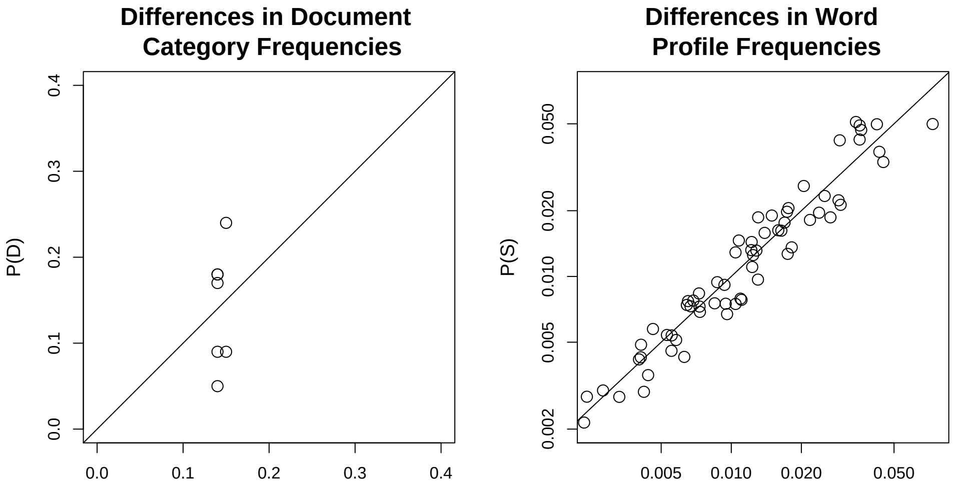

Figure 2: Differences between Labeled and Population Document Sets: For both $P(D)$ on the left and $P(S)$ on the right, the distributions differ considerably. This would make any of the standard supervised learning estimators massively biased. Our approach, as we show below, will be able to generalize from labeled to population sets despite these differences.

Despite the considerable differences between the labeled data set and the population, and the fact that even much smaller differences would bias standard approaches, our approach still produces accurate estimates. Figure 3 presents these results. The actual $P(D)$ is on the horizontal axis and the estimated version is on the vertical axis, with each of the seven circles representing one of the document frequency categories. Estimates that are accurate fall on the $45°$ line. In fact, the points are all huddled close to this equality line, with even the maximum distance from the line for any point being quite small.

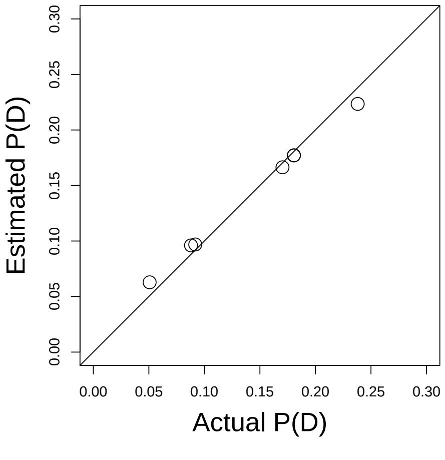

Figure 3: Comparing Nonparametric Estimates to the Truth: Despite the differences in $P(D)$ and $P(\boldsymbol{S})$ between the labeled and test sets shown in Figure 2, this figure shows how our nonparametric estimator remains unbiased.

7.2 Empirical Evidence

We now offer three direct out-of-sample tests of our nonparametric approach in different types of data. We begin with the 4,303 blog posts which mention George W. Bush.[9]

These posts include 47,726 unique words and 3,165 unique word stems among those appearing in more than 1% and fewer than 99% of the posts. We randomly divide the data set in half between the training set and test set and then randomly delete half (or 713) of the posts coded -2 or NB among those in the test set. Our test set therefore intentionally selects on (what would be considered, in standard supervised learning approaches) the dependent variable. This adjustment would create selection bias in the standard approach but, as we now show, leaves inferences from our approach approximately unbiased. The results from our nonparametric estimator appear in the top left graph in Figure 4. This graph plots one circle for each of the seven categories, with 95% confidence intervals appearing as a vertical line through it. Clearly most of the points are quite close to the 45 degree line, indicating approximately unbiased estimates, and all are within the 95% confidence intervals.

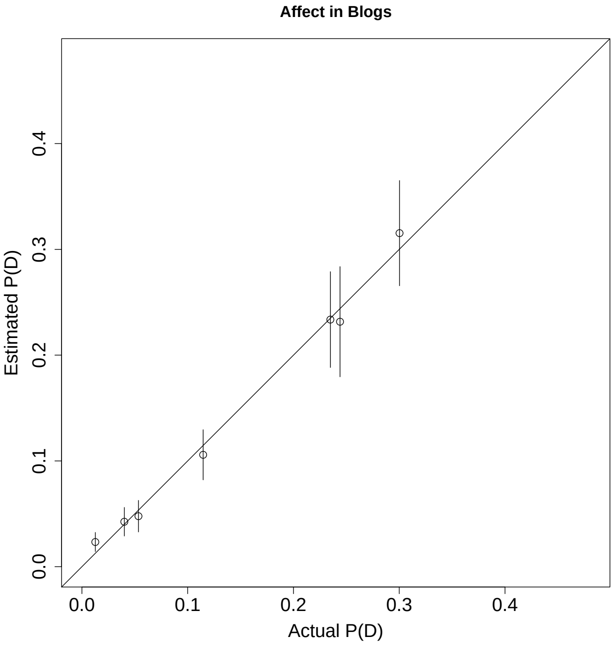

Figure 4: Out-of-sample Validation for the Nonparametric Method. Each plot gives the estimated document category frequencies (horizontally) by the actual frequencies (vertically), with 95% confidence intervals appearing as vertical lines. Estimates closer to the $45°$ line are more accurate.

Our second example is from a standard corpus of movie review ratings commonly used in the computer science literature to evaluate supervised learning methods (Pang and Lee, 2005).[10] The categorization is one, two, three, or four stars indicating the quality of the movie. To demonstrate

[9] As we show in Section 7.3, 4,303 hand coded documents is more than necessary, perhaps by a factor of 10. We started this project intending to use standard supervised learning methods, but despite numerous attempts over many months and numerous approaches, we found we could not produce unbiased estimates of our quantities of interest — at least when we maintained a highly rigorous evaluation standard. Even when we managed to to increase the percent correctly predicted, the degree of bias remained unacceptably high. So we kept hand coding while we tried technique after technique, until we thought of the ideas described herein. Needless to say, we plan to code fewer documents next time!

[10] We use the "scale data set v1.0," available at `http://www.cs.cornell.edu/people/pabo/movie-review-data`.

the method's effectiveness in cases where the test set is not drawn from the same population as the training set, we divided the 5,005 reviews into a training set of 2,672 reviews written by two authors and a training set of 2,333 reviews written by two other authors. The results using the nonparametric estimator appear in the left graph in Figure 5, and are again quite accurate. Deviations from the $45°$ line are due to slight violations of the assumption in Equation 9, perhaps due to category drift. Another example is from the Carnegie Mellon University Text Learning Group on university web sites, another standard computer science corpus. This project classified university webpages in 1997 as belonging in one of seven non-ordered categories (student, faculty, staff, department, course, project, or other). Using the four most common types of webpages, we extracted a data set of 2,050 webpages from Cornell and Wisconsin. We then used these data as a training set to estimate the distribution of webpages in a new corpus of 1,976 webpages from the Universities of Texas and Washington. The results appear in the center graph in Figure 5. Again, in a case where the training set and test set are drawn from different populations, our nonparametric method gives estimates that are all clustered near the $45°$ line, indicating accurate estimates.[11]

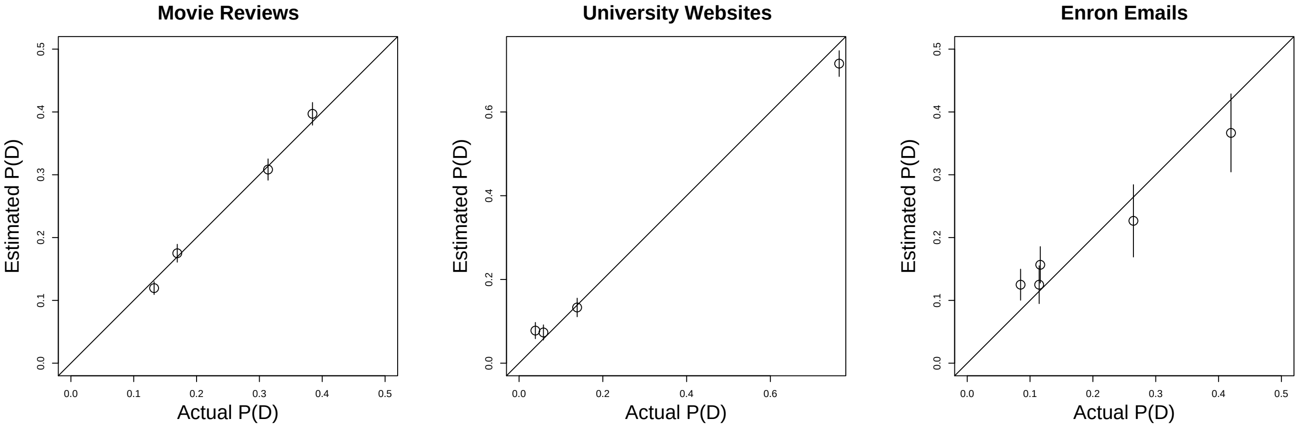

Figure 5: Additional Out-of-sample Validation. The left graph displays the accuracy of the nonparametric method in recovering the distribution of ordered categories for movie reviews. The center graph shows the same for a non-ordered categorization of university websites, and the right graph shows the same for a non-ordered categorization of emails sent by Enron employees. As before, 95% confidence intervals are represented by vertical lines, and estimates closer to the $45°$ line are more accurate.

Our final example comes from 1,726 emails sent by Enron employees (released during litigation) and classified into five non-ordered categories: company business, personal communications, logistic arrangements, employment arrangements, and document editing.[12] To make the task more difficult, we first created a skewed test set of 600 emails that was more uniformly distributed than the training set, with no category accounting for less than 12% or more than 39% of the observations. We then used the remaining 1,126 emails as a mutually exclusive training set where the comparable bounds were 4% and 50%. The results are quite accurate, especially given the paucity of information in many (short) emails, and are displayed in Figure 5. The method thus appears effective with shorter texts as well as longer texts.

[11] The original data appear at `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/`.

[12] For more on the data set, see `http://bailando.sims.berkeley.edu/enron_email.html`.

7.3 How Many Documents Need to be Hand Coded?

Any remaining bias in our estimator is primarily a function of the assumption in Equation 9 (see Section 9 for details). In contrast, efficiency, as well as confidence intervals and standard errors, are primarily a function of how many documents are hand coded. But how many is enough? Hand coding is expensive and time consuming and so we would want to limit its use as much as possible, subject to acceptable uncertainty intervals.

To study this question, we set aside bias by randomly sampling the labeled set directly from the population. For both our estimator (on the left) and the direct sampling estimator (on the right), Figure 6 plots the bias vertically by the number of hand coded documents in the labeled test set horizontally. Zero bias is indicated by a horizontal line.

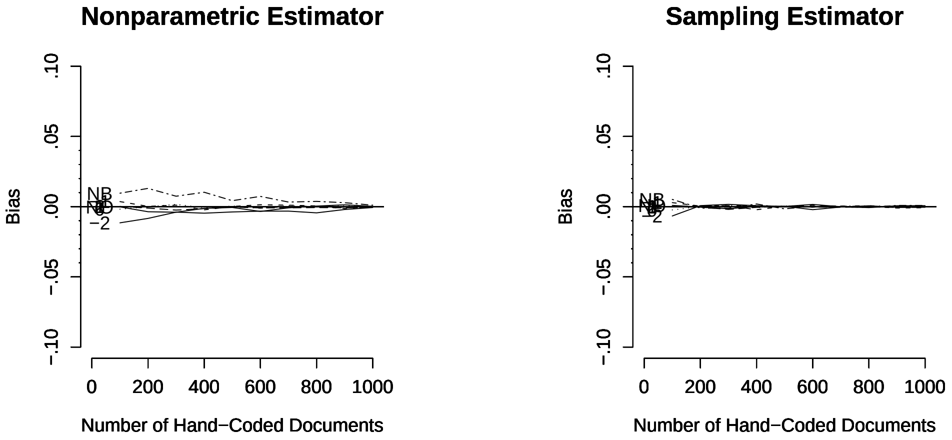

Figure 6: Bias by Number of Hand Coded Documents

In these data, a direct sampling approach is clearly optimal, and the right graph shows the absence of bias no matter how many documents are in the labeled set. Our estimator, in the left graph, is also approximately unbiased for the entire range of corpus sizes. That is, even for as few as 100 hand coded documents, both estimators are unbiased, and even the largest deviations of bias from zero is never much more than a single percentage point. The difference is that the dispersion around zero bias is slightly higher for our estimator than the error in direct sampling.

This pattern is easier to see in Figure 7 where we plot the root mean square error (RMSE) averaged across the categories vertically by the number of hand coded documents horizontally for our estimator (straight line) and the direct sampling estimator (dashed line). RMSE is lower for the direct estimator, of course, since this sample was drawn directly from the population and little computation is required, although the difference between the two is only about two tenths of a percentage point.

For our estimator, which will have considerably lower RMSE than the direct sampling when random sampling from the population is not possible, the RMSE drops quickly as the number of hand coded documents increase. Even the highest RMSE, with only 100 documents in the labeled set, is only slightly higher than 3 percentage points, which would be acceptable for many applications in the social sciences. (For example, most national surveys have a margin of error of at least 4 percentage points, even when assuming random sampling and excluding all other sources of error.) At about 500 documents, the advantage of more hand coding begins to suffer diminishing returns. In part this is because there is little more error to eliminate as our estimator then has an average RMSE of only about 1.5 percentage points.

The conclusion here is clear: coding more than about 500 documents to estimate a specific

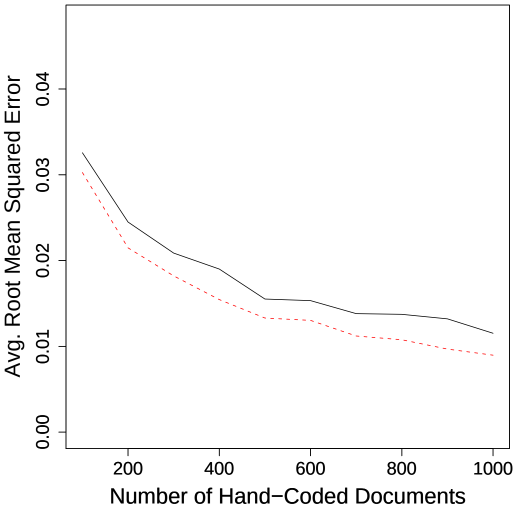

Figure 7: Average Root Mean Square Error by Number of Hand Coded Documents

quantity of interest is probably not necessary, unless one is interested in much more narrow confidence intervals than is common or in specific categories that happen to be rare. For some applications, as few as 100 documents may even be sufficient.

8 Correcting for Lack of Intercoder Reliability

As discussed in Appendix A, developing categories, training coders, and conducting large scale hand coding sets is often an error-prone task. Although scholars no longer concede that “The procedures and the categories used in content analysis cannot be standardized” (Waples, Berelson and Bradshaw, 1940), the difficulty of the task is widely recognized. Inter-coder reliability is measured in many different ways in the literature, but the rates tend to be lower with more categories and more theoretically interesting coding schemes. Reliability rates are not perfect in almost any study when documents are hand coded.

For example, we had at least two coders categorize each of 4,169 blog posts. In these data, our coders agreed on the classification of 66.5% of the blog posts; they agreed on 71.3% of blog posts among those when both coders agreed the post contained an opinion; and they agreed on 92% of the posts for an aggregated classification of negative, neutral, or positive opinions among posts with opinions. Table 1 gives more detailed information about these results. For any two coders, arbitrarily named 1 and 2, each row in the table gives the probability of coder 2’s classification given a particular classification d which coder 1 chose, $P(D_2|D_1 = d)$, with the marginal probability for coder 1 appearing in the last column, $P(D_1)$. The “misclassification” (or “confusion”) matrix in this table includes information from all combinations of observed ordered coder pairs. These numbers are comparable to other studies of hand coded data.

Unfortunately, “the classical supervised classification paradigm is based on the assumption that there are no errors in the true class labels” and “that the classes are well defined” (Hand, 2006, p.9). Indeed, almost all social science applications, including ours in Section 7, make this same dubious assumption. The problem may be due to “conceptual stretching” (Collier and Mahon, 1993; Sartori, 1970) or “concept drift” (Widmer and Kubat, 1996) that could in principle be fixed with a more disciplined study of the categories or coder training. Or it may be that no amount of

	-2	-1	0	1	2	NA	NB	$P(D_1)$
-2	**.70**	.10	.01	.01	.00	.02	.16	.28
-1	.33	**.25**	.04	.02	.01	.01	.35	.08
0	.13	.17	**.13**	.11	.05	.02	.40	.02
1	.07	.06	.08	**.20**	.25	.01	.34	.03
2	.03	.03	.03	.22	**.43**	.01	.25	.03
NA	.04	.01	.00	.00	.00	**.81**	.14	.12
NB	.10	.07	.02	.02	.02	.04	**.75**	.45

Table 1: This table presents conditional probabilities for coder 2's classification (in a set of column entries) given a code assigned by coder 1 (corresponding to a particular row), or $P(D_2|D_1)$. For instance, when coder 1 chooses category -2, coder 2 will choose the same category 70% of the time, category -1 10% of the time, category 0 1% of the time, and so on across the first row. This matrix is estimated from all 4,169 ordered coding pairs from five coders going in both directions. The final column denotes the marginal probability that coder 1 placed the blog in each category.

training could produce 100% reliability due to an inherent, irreducible uncertainty in representing human language in fixed categories. To some degree the latter is accurate, but a scientific approach to measurement means we must continually strive for better category definition, documentation, training, and evaluation.

Of course, no matter how long we try and how careful our procedures are, at some point we must stop and begin drawing conclusions from the data with whatever level of misclassification remains. Judging from the literature, this point is almost always reached prior to eliminating all risk of misclassification. For the lack of anything better to do, what virtually all scholars do when they stop is to pretend there exists no misclassification. We discuss the consequences of this procedure and a way to partially ameliorate the problem in this section.

Our idea is to adapt a technique called simulation-extrapolation (SIMEX) to the problem of imperfectly coded content analyses, which has not been done before. SIMEX is due to Cook and Stefanski (1994), turns out to be closely related to the jackknife (Stefanski and Cook, 1995), and has subsequently been applied in other areas (Carroll, Maca and Ruppert, 1999; Küchenohoff, Mwalili and Lassaffre, 2006). We first offer some intuition, then some formalization, and finally an empirical illustration.

Intuition For intuition, we illustrate our approach by an analogy to what might occur during a highly funded research project as a coding scheme becomes clearer, the coding rules improve, and coder training gets better. For clarity, imagine that through five successive rounds, we have different, more highly trained coders classifying the same set of documents with improved coding rules. If we do well, the results of each round will have higher rates of inter-coder reliability than the last. The final round will be best, but still not perfect. If we could continue this process indefinitely, we might succeed in banishing all misclassification, but this is typically infeasible.

Now suppose our estimate of the percent of documents in category 2 is 5% in the first round, 11% in the second, 14% in the third, 19% in the fourth, and 23% in the last round. The question, then, is what to do once we observe these results. The procedures used by all previously published content analyses would have us use 23% as the best estimate of the proportion in category 2. This is not an unreasonable approach, but our point is that it appears to leave some information on the table and thus might be improved on. In particular, if the proportion of documents in category 2 is increasing steadily as the level of inter-coder reliability at each round improves, then we might reasonably extrapolate this proportion to the point where inter-coder agreement is perfect.

We might thus conclude that the true proportion in category 2 is actually somewhat larger than 23%. We might even formalize this idea by building some type of regression model to predict the category 2 proportion with the level of inter-coder reliability and extrapolate to the unobserved point where reliability is perfect. Since this procedure involves extrapolation, it is inherently model dependent and so uncertainty from its inferences will exceed the nominal standard errors and confidence intervals (King and Zeng, 2006). However, a crucial point is that even using the figure from the final round and doing no subsequent processing still involves an extrapolation; it is just that the extrapolation ignores the information from previous rounds of coding. So using 23% as our estimate and ignoring this idea is no safer.

Formalization Since firing one's coders after each round of training makes learning and improving less likely, and in any event does not happen, we make use of the misclassifications estimated from a single round of coding with more than one coder, simulate what would have happened to the document category proportions if there were even lower levels of inter-coder reliability, and extrapolate back to the point of no misclassification.

To formalize this SIMEX procedure, we begin with our estimation method, which would give statistically consistent answers if it were applied to data with no misclassification. The same method applied to error-prone data is presumably biased. However, in this problem, the type of misclassification is easy to characterize, as we do in Table 1. Then we follow three steps: (1) Take each observed data point D_i in the labeled set and simulate M error-inflated pseudo-data points, using the misclassification matrix in Table 1. We do this by drawing M values of $\tilde{D}_i$ from the probability density $P(\tilde{D}_i|D_i)$ (given the observed data point D_i) which appears in the corresponding row of the table. This step creates M simulated data sets with twice the amount of measurement error, of the same type as in our observed data, to these pseudo-data. We then repeat this procedure starting with these pseudo-data to produce M pseudo-data sets with three time the measurement error as in the original data. Then again with four times the amount of measurement error, etc. (2) We apply our estimator to each of the simulated pseudo-data sets and average over the M results for each level of added error. This leads to a sequence of averaged results from each of the pseudo-estimators, with a different level of inter-coder reliability. (3) We transform these data using the multivariate logistic transformation, and then (4) fit a relationship between the transformed average proportion of observations estimated to be in each category from the error-inflated pseudo-data sets and the amount of added error in each. We then (5) extrapolate back to the unobserved point of zero measurement error, and transform the results so that they are again constrained to fall on the $(0, 1)$ interval.

Illustration Figure 8 gives an example of this procedure for one category from our blogs data. The vertical axis of this graph is the proportion of observations in category NB. The horizontal axis, labeled α, gives the number of additional units of misclassification error we have added to the original data, with the observed data at value 0. The estimate of the element of $P(D = \{\text{NB}\})$ from the original data (corresponding to the last round of coding from the example in the previous paragraph) is denoted with a diamond above the value of zero. A value of α of 1 means that the original data went through the misclassification matrix in Table 1 once; 2 means twice, etc. Some noninteger values are also included. In the application, it seems likely that the proportion of documents we would have estimated to be in category NB, if our coders had perfect rates of inter-coder reliability, would be higher than the proportion from our actual observed data.

All applications begin with the point estimated from the observed data at zero (marked by a diamond in the figure), and extrapolate it over to the horizontal axis value of -1, which denotes the data with no misclassification error. The implicit extrapolation used in almost all prior content analysis research occurs by effectively drawing a flat line from the diamond to the vertical axis

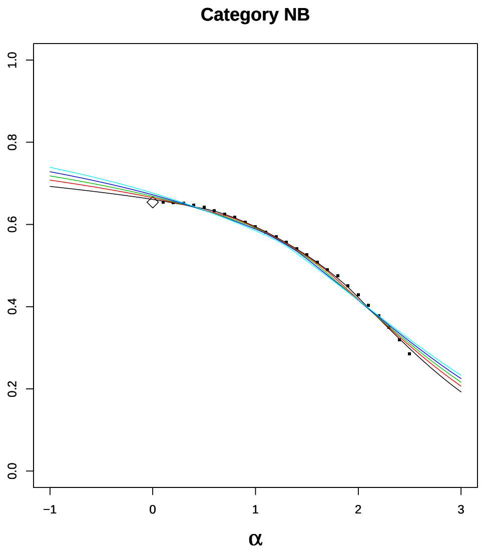

Figure 8: SIMEX Analysis of the Proportion of documents in Category NB (not a blog). The estimate from the observed data appears above 0 marked with a diamond; other points are simulated. The goal is to decide on the proportion in category NB at a horizontal axis value of -1.

on the left; this flat line is used regardless of what information is available. This procedure is not always wrong but it comes with no evidence that it is right and obviously offers no flexibility when new information arises.

The question is whether there might be sufficient information in the simulated error-inflated data to extrapolate better than using the existing flat line extrapolation procedure. In most cases, we believe there is. In the example in Figure 8, estimates from these error-inflated data also appear, as well as several alternative (LOESS-based) models used to form possible extrapolations. The result is model dependent by nature, but the same is the case whether we use or ignore the information from the simulated data.

Figure 9 presents analogous results from the remaining six document categories. In each case, there appears to be some additional information that may be useful in extrapolating the true proportion of documents to the left of the curve. Some of the uncertainty in extrapolation is illustrated in the graphs via separate lines, each from a different method used to extrapolate, but of course numerous other models could have been used instead. The key point of this section is that extrapolation is necessary whether or not this SIMEX procedure is used. The only other choice is to go back to trying to improve the categories, coding rules, and coder training.[13]

9 What Can Go Wrong?

We now discuss five problems that can arise with our methods. If they do arise, and steps are not taken to avoid or ameliorate them, they can cause our estimator to be biased or inefficient. We also discuss what to do to try to ameliorate these problems.

A key issue is the assumption in Equation 9 that $P(\boldsymbol{S}|D)$ is the same in the labeled and population document sets. So if we are studying documents over a long time period, where the

[13] If we have an inexpensive hand-coded data set and an expensive gold standard data set, even for some smaller proportion of observations, we could correct not only for the amount of measurement error but also for the specific type (see Küchenohoff, Mwalili and Lassaffre, 2006). Unfortunately, this type of information is both rare and would be unlikely to take us to the situation with fully corrected classifications.

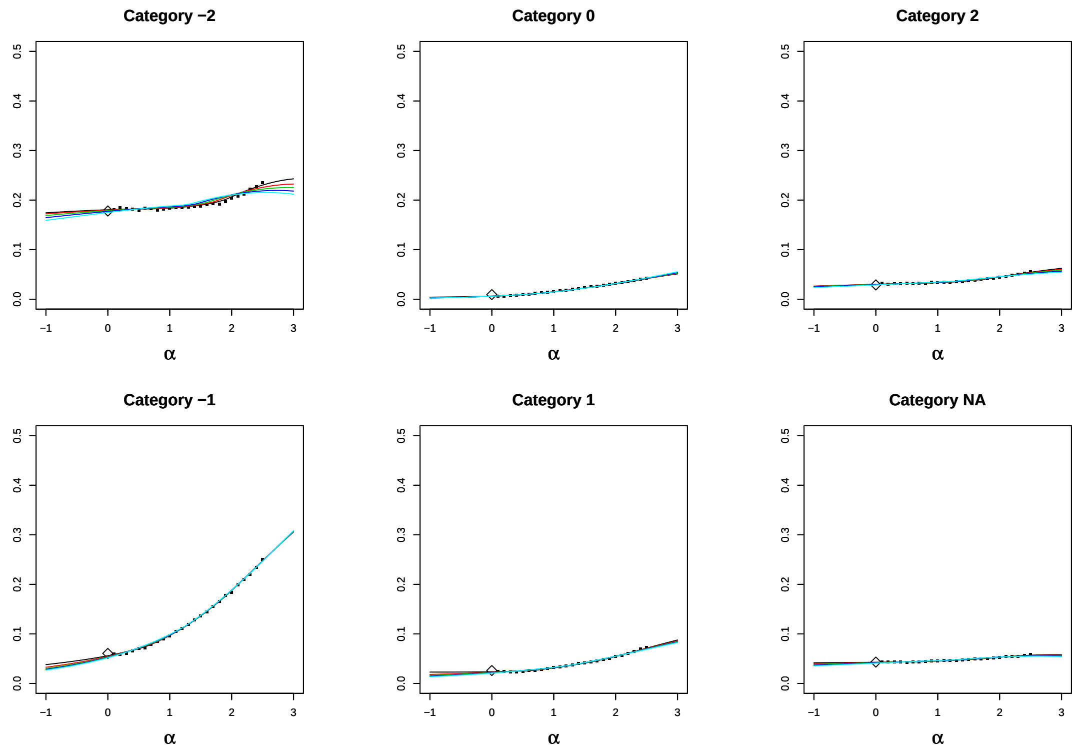

Figure 9: SIMEX Analysis of the proportion of documents in the remaining six categories, a continuation of Figure 8.

language used to characterize certain categories is likely to change, it would not be advisable to select the labeled test set only from the start of the period. Checking whether this assumption holds is not difficult and merely requires hand coding some additional documents closer to the quantity presently being estimated and using them as a validation test set. If the data are collected over time, one can either hand code several data sets from different time periods or gradually add hand coded documents collected over time. In our running example, we are attempting to track the opinion of those with opinions over a single presidential campaign. As such, only one hand-coded data set at the start may be sufficient, but we have tested this assumption, and will continue to do so, by periodically hand coding small numbers of blogs over time.[14]

Second, as King and Lu (2007) describe, each category of D should be defined so as not to be too heterogeneous. In our work, if a category is highly heterogeneous, then the set of word profiles that could be used to characterize the language used will be too large to describe all the documents in that category. This is easy to see when you try to describe the category and find yourself using many different examples. This problem seems more likely to occur for residual or catch-all categories. Imagine how you would describe to someone else category “NB” (not a blog) in our data. This is difficult since there are innumerable types of web sites that are not blogs, each with potentially very different language; yet this category was essential since our blog search algorithm was not perfect. We do find slightly more bias in estimating category NB than the others

[14] To generate a clear example of where this assumption is violated, we divided our labeled blogs into training and test sets and then further divided the test set into subsets based on the sophistication of the language. We did so using the Flesch-Kincaid grade level score, which uses words per sentence and syllables per word to provide an indication of the grade level needed to read a text. We then tried to estimate the document category frequencies from a labeled set that made no such distinctions. Since the language sophistication is computed directly from the text of the documents, the assumption in Equation 9 was then violated and our estimates were biased as a result.

in our categorization. The small but noticeable bias in this category is apparent if you look closely at the top line on the left in Figure 6.

Third, our approach requires the choice of the number of word stems to use in each randomly chosen subset, when estimating $P(D)$. While choosing the number of random subsets is easy (the more the better, and so like any simulation method should be chosen based on available computer time and the precision needed), the number of word stems to use in each random subset must be chosen more carefully. Choosing too few or too many will leave $P(\boldsymbol{S})$ and $P(\boldsymbol{S}|D)$ too sparse or too short and may result in attenuation bias due to measurement error in $P(\boldsymbol{S}|D)$, which serve as the "explanatory variables" in the estimation equation. To make this choice in practice, we use standard cross-validation techniques, such as by dividing the labeled set into a training and test set and then see what works in those data. (We recommend making this division randomly to ensure that this auxiliary study is not confounded by potential violations of the assumption in Equation 9.) The algorithm is not very sensitive to this choice, and so there is typically a range of values that work well. In practice, the number of word stems to choose to avoid sparseness bias mainly seems to be a function of the number of unique word stems in the documents. Although one can make the wrong choice, and making the right choice may take some effort, fixing any problem that may arise via these types of cross-validation tests is not difficult.

Fourth, we require a reasonable number of documents in each category of D to be hand coded. Although we studied the efficiency of our procedure as a function of the number of hand coded documents in Section 7.3, these results would not hold if by chance some categories had very few hand coded documents and we cared about small differences in the proportions in these population categories. This makes sense, of course, since the method requires examples from which to generalize. Discovering too few examples for one or more categories can be dealt with in several ways. Most commonly, one can alter the definition of the categories, or can change the coding rules.

However, even if examples of some categories are rare, they may be sufficiently well represented in the much larger population set to be of interest to social scientists. To deal with situations like this, we would need to find more examples from these relatively rare categories. Doing so by merely increasing the size of the hand coded data set would usually not be feasible and in any event would be wasteful given that we would wind up with many more coded documents in the more prevalent categories. Still, it may be possible to use available metadata to find the needed documents with higher probability. In our blogs data, we could find blog posts of certain types via links from other posts or from popular directories of certain types of blogs. Fortunately, the labeled set is assumed to be generated conditional on the categories, and so no bias is induced if we add extra examples of certain categories. In other words, we are already assuming that $P(D)$ may differ in the labeled and population sets and so selecting on D to over-represent some categories causes no difficulties with our procedure (cf. King and Lowe, 2003).

Finally, and most importantly, our procedure cannot work without access to reliable information. This requires that the original documents contain the information needed, the hand codings are reliable enough to extract the information from the documents, and the abstract quantitative summary of the document (in $\boldsymbol{S}$) is a sufficiently accurate representation and enough to estimate the quantities of interest. Each of these steps require careful study. Documents that do not contain the information needed cannot be used to estimate quantities of interest. If humans cannot code these documents into well-defined categories with some reasonable level of reliability, then automated procedures are unlikely to succeed at the same task. And of course many choices are available in producing abstract numerical summaries of written text documents.

Throughout all these potential problems, the best approach seems to be the radically empirical procedure suggested in the supervised learning literature: If the procedure you choose works, it works; if it doesn't, it doesn't. And so be sure to verify that your procedures work, subdividing your labeled set into training and (truly out of sample) test sets and directly testing hypotheses about the success of the procedure. Ideally, this should then be repeated with different types of labeled test sets. The more we make ourselves vulnerable to being wrong, using rigorous scientific procedures, the more we learn. Fortunately, the tools we make available here would seem to make it possible to learn enough to produce a reliable procedure in many applications.

Relatedly, standard errors and confidence intervals take a very different role in this type of research than the typical observational social science work. Unlike many social science problems, if your uncertainty is too large here, you merely need to hand code additional documents. In fact, sequential sampling is perfectly appropriate: After finding a valid categorization scheme, hand code say 100 documents and compute the quantities of interest and their confidence intervals or standard errors. If these estimates indicate more uncertainty than desired, then hand code more documents, add them to the first set, and reestimate. No bias will be induced by this sequential sampling plan so long as the selection of additional documents follows some appropriate random procedure unrelated to the quantities estimated.

10 Concluding Remarks

Our approach to reading text and extracting a specific type of systematic social scientific information from it requires no modeling assumptions, no modeling choices, and no complicated statistical approaches, and lets the social scientist pose the theoretical question to be answered. It also requires far less work than projects based entirely on hand coding and can be done both fast and in real time. Individual-level classification is not a result of this method, and so it is not useful for all tasks, but numerous quantities of interest, from separate subdivisions of the population or different populations, can be estimated. Our method does require careful efforts to properly define categories and to hand code a small sample of texts.

We believe that the methods offered here open up a range of analyses that may not have been previously feasible. With the explosion of numerous types and huge quantities of text available to researchers on the web and elsewhere, we hope social scientists will begin to use these methods, and develop others, to harvest this new information and to improve our knowledge of the political, social, cultural, and economic worlds.

A Coding Schemes

The most difficult part of any content analysis project, using either an entirely hand-coded approach or some supervised learning method, including ours, is producing an acceptable categorization scheme. The difficulty is often surprising and frustrating to those who come anew to content analysis projects, but obvious once you try it yourself. The problem (and opportunity) is that human language and reasoning admits to an extraordinarily large and complicated set of possible expressed opinions, and no theory exists that can reliably predict what categorization scheme will work for any given set of documents ex ante.

Unfortunately, these problems are not often discussed in sufficient detail in published research. Although it appears to be the same in most areas of application, Kolbe and Burnett (1991) summarize an extensive review of content analyses in consumer research by writing “Most factors pertaining to objectivity were either unreported or unattended by authors. Problems with reliability reporting. . . were also present.” Our reading of the social science content analysis literature, as well as what we learned from personal consultations with experienced researchers in the area, indicates that a great deal of work typically goes into improving reliability. However, this work is not often discussed in published work, success remains at best uneven, and intercoder reliability statistics go unreported (or not fully reported) in about half of all published content analysis work.

We thus pause briefly to offer some suggestions from our experiences.

To begin, we summarize the basic rules of coding: (1) Find a categorization scheme that encodes the social science question being asked in categories that are mutually exclusive (so each document goes into only one category) and exhaustive (so all documents are classified). The categories may be purely nominal, ordered, partially ordered, unidimensional, multidimensional, or mixed. If the categories of interest are not mutually exclusive, researchers should determine subsets that are mutually exclusive (such as by defining categories for "both" or neither"), or else can categorize texts on multiple dimensions.[15] (2) Produce a coding manual clear enough so that coders can be trained, at least in principle, by looking only at the manual and not talking to the researchers (thus ensuring that the research procedures are fully public, do not require the author's involvement, and so are replicable at least in principle). (3) Measure the extent to which different coders can code the same documents in the same ways (i.e., measure inter-coder reliability). And (4) check for validity, ideally by comparison to some external gold standard, or more commonly in practice by the researcher reading and checking that the categories and codings reflect the theoretical concepts of interest. Of course, one cannot have validity without reliability and so much of the work in producing a coding scheme involves iterating between category definition and inter-coder reliability checks.[16]

As in most areas of science it is too easy to fool oneself, in this case that the four rules are satisfied. In particular, checking inter-coder reliability is time-consuming and expensive, and ensuring that your communications with coders are formalized in a set of written documents sometimes feels like an unnecessary hurdle in the way of progress. However, rigorous evaluation — large numbers of documents coded by two or more coders who only read a set of coding rules and do not interact with each other while coding — has no substitutes. Similarly, the various compromises that are made, such as having a third coder resolve discrepancies, having two coders who disagree talk out their differences, or communicating the coding rules to coders in long training sessions or conversations, may often be reasonable after a coding scheme is established, but they can make rigorous evaluation impossible.

The fundamental difficulty in meeting these rules is that categorization schemes that seem theoretically appropriate before confronting the data often turn out to have low inter-coder reliability. Studying the exceptions then quickly reveals problems with the measurement "theory." Adjustments can then go in two directions, perhaps at the same time. One is to further articulate the categories, and the other is to simplify, usually by combining categories. The former may be theoretically more attractive, but it imposes even more burdens on coders, and so can lead to lower levels of inter-coder reliability.

We illustrate these difficulties (and others) through a few of our attempts to find appropriate coding schemes for blogs. For one simple example, we tried at one point to use a dichotomous coding rule for whether a blog post was about the *policies* of President Bush and his administration or his personal *character*. Judging from the political science literature, this ought to be easy. It is a standard theoretical distinction in a large body of research, and is represented in many of our survey questions, theoretical discussions, and empirical analyses (Marcus, 1988, 2006; Ragsdale, 1991; Bishin, Stevens and Wilson, 2006). But all this only means that the distinction is logically consistent and of theoretical interest; it does not mean that ordinary people express themselves the way these and other creative professional political scientists conceptualize the world. Indeed, we found that no matter how precise we made our coding rules, and how we trained our coders, our

[15] The coding scheme in Section 2 illustrates a number of these features. One possibility not included is categories with higher dimensional orderings, such as affect by partisanship, in addition to categories like NA and NB. One can also code the same documents in multiple parallel categorization schemes.

[16] This iterative process between theory and data further blurs the lines between the categories of supervised vs. unsupervised learning methods. Both approaches discover categorization schemes from the data to at least some degree.

inter-coder reliability rates in classifying according to this rule were embarrassingly low.

For example, consider the following two excerpts, each from a separate blog post, which clearly do not fit the policy/character distinction:

> ”. . . What do I see in Bin Laden’s words? I see a very desperate criminal trying in every way he knows to get us to agree to stop pushing him. I see that we are winning and he knows it, thus he will try to hit us in the spots that he thinks are weak so that we will give up. That means he will find the lies to which he discerns that Americans are susceptible and tell them to us. I am glad once again that Bush is president and not someone who would back down in front of terrorist threats.”[17]
>
> “In spite of market and administration hype, the economy is on the decline. There are no positive trends. The latest unemployment statistics are deliberate, numerical deceptions. Using the workforce participation rate used in January of 2001, the current unemployment rate is actually 7.2%, not 5.2% as deceptively stated by the Bush administration.”[18]

We tried to respond to exceptions like these with various common coding tricks. For example, we instructed coders to identify the primary thrust of the criticism. Then we told them that any criticism of policy should be classified as policy even if it also mentions character attributes. We tried telling them that references only to specific policy actions should be coded as policy. We also experimented with coding rules where posts which do not reference specific policies are to be coded as character. We tried to further articulate the categories, by including a “both” category, in addition to “policy” and “character,” to deal explicitly with ambiguous posts, but inter-coder reliability remained low. It turns out that deciding when a post was “both” was itself highly ambiguous, and that category turned out to have almost zero inter-coder reliability! No coding scheme for this distinction came close. Thus, we eventually concluded that a categorization scheme based on this standard political science distinction was not feasible to use for categorizing political commentary by ordinary Americans. We also tried more detailed categorization schemes, including a 25 category scheme developed from the list of 18 instrumental values offered by Rokeach (1973), but this and other schemes failed too. When the distance spanned by theory and the empirical world is this large, it pays to modify one’s theory.

B Example Blog Posts In Our Categories

Below are excerpts from blog posts about Senator Hillary Clinton that exemplify each of our five ordered sentiment categories.

-2 McCain is most likely going to be a candidate for prez, and he’ll most likely be running against Horseface Hillary. So now he must discredit her all he can. . . . I sure as hell wouldn’t vote for McCain, unless the only options were him and Hillary. (Hillary Is An Idiot Blog, 10/11/06)

-1 I am one of those who is extremely disappointed in John McCain. Altho I did disagree with alot of his positions, I respected the fact that he was a straight talker, so that at least I would listen to what he had to say and ACTUALLY consider his position and any merits it might contain—and THAT is what is needed if we are to get out of this partisan BS mess. However he has become such a suck-up lackey lap dog of the Bush whitehouse. I have just lost any

[17] `http://floa.blogspot.com/2006_01_01_archive.html`

[18] `http://unlawflcombatnt.blogspot.com/2006/11/economy-updates.html`

and all respect for the man, especially after what Bush and Co did to him in the primaries, that he could even offer any support to that slimeball only indicates the depth that he has sunk, and the complete loss of any dignity he might have once owned. I don't know why or how Hillary got into this discussion. I would rather not see her run just because I believe it WILL energize the far right "never vote FOR a democrat—would rather elect a bumbling idiot (proven in 2000 and 2004) than ANY democrat" type of voter to come to the polls. (Watchblog, 8/23/06)

0 Sen. Hillary Rodham Clinton apologized for joking that Mahatma Gandhi used to run a gas station in St. Louis, saying it was a lame attempt at humor.' Needless to say, there was criticism. But, was this really anything that we haven't all done? Of course we have. Not one of us is blameless. The lesson here is not that Hillary is a bigot. She's not. The lesson is that people say stupid stuff. If they genuinely apologize... accept it. Then, move on. (QandO Blog, 1/6/04)

1 Is it possible that Hillary won't be the one? As I've heard more than one fiercely loyal Democrat say (and I echo): I hope she doesn't get the nomination, but if she does I'll be 100% behind her. I really like her and would be happy to have her as my President, but I think that she's so polarizing that she can't possibly win. Of course people said that her husband didn't have a chance either. (I'm Not Crunchy Blog, 5/26/06)

2 Wow. Just 2 points between Hillary and the crazy John McCain? Really? Hellz Yeah. That rocks. All the nay sayers need to shut up. The Clintons' know what is required to win and they are willing to do it. This is why I love them both. There is absolutely nothing that can stop a Clinton/Clark campaign. All they have to do, again, is hold Kerry's states. That shouldn't be too much of a problem. And pick up OH, FL or AR. This can be done, people. (BRock NYLA Blog, 9/27/06)

References

Adamic, L.A. and N. Glance. 2005. "The political blogosphere and the 2004 US election: divided they blog." *Proceedings of the 3rd international workshop on Link discovery* pp. 36–43.

Benoit, Kenneth and Michael Laver. 2003. "Estimating Irish party policy positions using computer wordscoring: the 2002 election - a research note." *Irish Political Studies* 18(1):97–107.

Berelson, B. and S. de Grazia. 1947. "Detecting Collaboration in Propaganda." *Public Opinion Quarterly* 11(2):244–253.

Bishin, Benjamin, Daniel Stevens and Christian Wilson. 2006. "Character Counts: Honesty and Fairness in Election 2000." *Public Opinion Quarterly* 70(2):235–248.

Carr, David. 2007. "24-Hour Newspaper People." *New York Times* (15 January).

Carroll, Raymond J., Jeffrey D. Maca and David Ruppert. 1999. "Nonparametric regression in the presence of measurement error." *Biometrika* 86(541-554):3.

Cavnar, W.B. and J.M. Trenkle. 1994. "N-Gram-Based Text Categorization." *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrival* pp. 161–175.

Collier, David and James E. Mahon, Jr. 1993. "Conceptual 'Stretching' Revisited: Adapting Categories in Compartive Analysis." *American Political Science Review* 87(4, December):845–855.

Cook, J. and L. Stefanski. 1994. "Simulation-extrapolation estimation in parametric measurement error models." *Journal of the American Statistical Asociation* 89:1314–1328.

Das, Sanjiv R. and Mike Y. Chen. 2001. "Yahoo! for Amazon: Opinion Extraction from Small Talk on the Web." Sanjiv Das, Department of Finance Santa Clara University.

Drezner, Daniel W. and Henry Farrell. 2004. “The Power and Politics of Blogs.” American Political Science Association, Chicago, Illinois.

Gamson, William A. 1992. *Talking Politics*. New York, NY: Cambridge University Press.

Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco and Judith L. Weddle. 1994. “Machine Coding of Event Data Using Regional and International Sources.” *International Studies Quarterly* 38(1):91–119.

Grindle, Merilee S. 2005. “Going Local: Decentralization, Democratization, and the Pomise of Good Governance.” Kennedy School of Government, Harvard University.

Hand, David J. 2006. “Classifier Technology and the Illusion of Progress.” *Statistical Science* 21(1):1–14.

Hindman, Matthew, Kostas Tsioutsiouliklis and Judy A. Johnson. 2003. “Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web.” Midwest Political Science Association, Chicago, Illinois.

Huckfeldt, R. Robert and John Sprague. 1995. *Citizens, Politics, and Social Communication*. New York, NY: Cambridge University Press.

Jones, Bryan D. and Frank R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago, IL: University of Chicago Press.

King, Gary and Langche Zeng. 2006. “The Dangers of Extreme Counterfactuals.” *Political Analysis* 14(2):131–159. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

King, Gary and Will Lowe. 2003. “An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design.” *International Organization* 57(3, July):617–642. http://gking.harvard.edu/files/abs/infoex-abs.shtml.

King, Gary and Ying Lu. 2007. “Verbal Autopsy Methods with Multiple Causes of Death.”. http://gking.harvard.edu/files/abs/vamc-abs.shtml.

Kolari, Pranam, Tim Finin and Anupam Joshi. 2006. “SVMs for the Blogosphere: Blog Identification and Splog Detection.” American Association for Artificial Intelligence Spring Symposium on Computational Approaches to Analyzing Weblogs.

Kolbe, R.H. and M.S. Burnett. 1991. “Content-Analysis Research: An Examination of Applications with Directives for Improving Research Reliability and Objectivity.” *The Journal of Consumer Research* 18(2):243–250.

Krippendorff, D.K. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Küchenohoff, Helmut, Samuel M. Mwalili and Emmanuel Lassaffre. 2006. “A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX.” *Biometrics* 62(March):85–96.

Laver, Michael, Kenneth Benoit and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *American Political Science Review* 97(2):311–331.

Lenhart, Amanda and Susannah Fox. 2006. Bloggers: A Portrait of the Internet’s New Storytellers. Technical report Pew Internet and American Life Project. http://207.21.232.103/pdfs/PIP

Levy, P.S. and E. H. Kass. 1970. “A three population model for sequential screening for Bacteriuria.” *American Journal of Epidemiology* 91:148–154.

Lyman, Peter and Hal R. Varian. 2003. How much information 2003. Technical report University of California. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

Marcus, George E. 1988. “The Structure of Emotional Response: The 1984 Presidential Candidates.” *American Political Science Review* 82(3):737–761.

Marcus, George E. 2006. “Emotions in Politics.” *Annual Review of Political Science* 3:221–50.

Mutz, Diana C. 1998. *Impersonal Influence: How Perceptions of Mass Collectives Affect Political Attitudes*. New York, NY: Cambridge University Press.

Neuendorf, K.A. 2002. *The Content Analysis Guidebook*. Sage Publications.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*. pp. 115–124.

Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. 2002. “Thumbs Up? Sentiment Classification using Machine Learning Techniques.” *Proceedings of the Conference on Empirical Methods in Natural Language Processing* pp. 79–86.

Porter, M. F. 1980. “An algorithm for suffix stripping.” *Program* 14(3):130–137.

Purpura, Stephen and Dustin Hillard. 2006. “Automated Classification of Congressional Legislation.” *Proceedings of the International Conference on Digital Government Research* .

Quinn, K.M., B.L. Monroe, M. Colaresi, M.H. Crespin and D.R. Radev. 2006. “An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th US Senate.” Annual Meetings of the Society for Political Methodology.

Ragsdale, Lyn. 1991. “Strong Feelings: Emotional Responses to Presidents.” *Political Behavior* 13(1):33–65.

Rokeach, Milton. 1973. *The Nature of Human Values*. New York: Free Press.

Sartori, Giovanni. 1970. “Concept Misformation in Comparative Politics.” *American Political Science Review* 64(4, December):1033–1053.

Simon, Adam F. and Michael Xeons. 2004. “Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis.” *Political Analysis* 12(1):63–75.

Stefanski, L.A. and J.R. Cook. 1995. “Stimulation-Extrapolation: The Measurement Error Jackknife.” *Journal of the American Statistical Association* 90(432, December):1247–1256.

Thomas, Matt, Bo Pang and Lillian Lee. 2006. “Get out the vote: Determining support or opposition from Congressional floor-debate transcripts.” *Proceedings of EMNLP* pp. 327–335. http://www.cs.cornell.edu/home/llee/papers/tpl-convote.home.html.

Verba, Sidney, Kay Lehman Schlozman and Henry E. Brady. 1995. *Voice and Equality: Civic Volunteerism in American Politics*. Cambridge, MA: Harvard University Press.

Waples, D., B. Berelson and F.R. Bradshaw. 1940. *What Reading Does to People: A Summary of Evidence on the Social Effects of Reading and a Statement of Problems for Research*. The University of Chicago Press.

Widmer, G. and M. Kubat. 1996. “Learning in the presence of concept drift and hidden contexts.” *Machine Learning* 23(1):69–101.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York, NY: Cambridge University Press.