Using Text Categorization to Measure the Impact of News on Public Opinion

James Allan, Alvaro Bolivar, and Courtney Wade
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003 USA
{allan,alvarob,cwade}@cs.umass.edu

Abstract

Public opinion researchers have investigated the question of whether news reporting affects the results of opinion polls that ask for the most important problem facing the nation. Those studies have generally used small collections of news in order to carry out their manual analysis of the relationships.

In this study, we used about 100 opinion polls gathered by several organizations from 1990 through 1998. The polling questions were manually combined into broad categories such as economics, health, and so on. Rather than manually categorize every news story, we classified a small number by hand and then used the BoosTexter text classification system to categorize over two million news stories in the same nine years.

We describe the process of cleaning our data to make it useful and the method for categorization. We then describe the results of our analysis of the relationship between polling and reporting. We show that more than half of the categories are strongly affected by reporting, but that the relationships are often much more complex than expected. Even though we were able to create the news corpus in substantially less time, our results are consistent with that of public opinion researchers.

1 Introduction

Does the public believe that gun control is an important issue because the news media cover it so extensively? Or do the media include so many stories about gun control because the public views it as a major issue? Or is there perhaps a more complex or subtle relationship between reporting and opinion?

Research on whether there exists such a relationship between media coverage of news events and issues and public opinion about these events and issues has traditionally been the domain of social scientists. They employ research methods such as survey research, content analysis, and statistical analysis. The process requires manually classifying opinion polls and news stories during the same period and then carrying out some form of regression analysis to look for relationships. The reliability of their results is determined to some degree by the amount of data that they can collect.

In this work, we introduce automatic text classification techniques to the study of this topic, greatly expanding the amount of data available for regression, making it possible to do a more complete time series analysis. Rather than manually classifying hundreds or perhaps several thousand news stories, we will classify over two million stories and perform regression to explore relationships between opinion and reporting. All classification is an errorful process, but automatic approaches are particularly suspect because there is little to no human oversight. We will show that our results are comparable to those found in several social science studies, strongly suggesting that our approach could be used to expand the data available for a range of social science analysis tasks.

In the next section we discuss some work related to that discussed in this study. Our process first requires gathering polling data and making it usable for our purposes, steps described in Section 3. Then in Section 4 we discuss how we gathered over two million news stories and categorized them into the same topics used for the polling data. In Section 5 we describe the regression analysis that we carried out and discuss the results. Consistent with studies within social science, we find that reporting on a topic affects opinion on the same topic, but that the reverse is not particularly true. We conclude in Section 6.

2 Related work

Most research in this area has been by political scientists who are interested in factors that predict the "public agenda" or "policy preferences," as measured by public opinion polls, regarding foreign and domestic policy issues. Most of the prior research has found that news coverage of certain topics does have an effect on public opinion. Jordan [3] and Page et al.[7] used pairs of identical survey questions on different topics and manually coded news stories beginning two months prior to the first survey to look at this issue. Jordan used a New York Times corpus, whereas Page et al. used television news. However, both found that news coverage accounted for at least some of the change in public opinion.

Two other studies exploring the relationship between news and public opinion, focused on specific topics. One attempted to understand the influence of different types of TV news coverage on public opinion on foreign policy issues [4], while another again looked at the relationship between TV news and public opinion, this time focusing on inflation, unemployment, and energy issues [1]. This latter study is of particular relevance because, like our study, the authors used surveys about the most important problem facing the nation to gauge public opinion. They also looked at the influence of poll results on news coverage. Their final regression models were similar to ours, except that they included "real world" factors such as the change in the unemployment rate as independent variables. Results indicated that there was a relationship in both directions—i.e., news affected polls *and* polls affected news.

While many of these papers do not provide details of the data collection and news coding process, the data used by Page et al. [7] and by Jordan and Page [4] took 29 research assistants 10,000 person-hours to compile.

3 Polling data

Since the inception of the Gallup Poll in 1935, the Gallup Organization and other polling organizations have intermittently conducted surveys that ask the open-ended question "What is the most important problem facing this country today?" A series of 102 public opinion polls[1] that included this question from 1990 through 1998 serve as the measure of public opinion for this study. The polls come from these sources:

Source	Count
Gallup	31
CBS/New York Times	42
ABC/Washington Post	19
Princeton Survey Research Associates	9
The Tarrance Group, K.R.C. Research	1
TOTAL	102

One of the items of information depicted in the complex chart of Figure 3 is the dates of all polls. Along the top of the chart are black diamonds and grey triangles, the former representing polls from the Gallup Organization and the latter representing all other polls. (The differentiation is historical: our initial polling data was exclusively Gallup polls. We were fortunate enough to find additional polls to fill out the data.)

[1] One of the Gallup polls (July 19-22, 1990) had percentages that only added up to 88. We contacted the Roper Center to acquire corrected numbers for that poll.

Because the "most important problem" question is open-ended, the polling organizations group responses into categories for reporting[2]. For the period 1990 to 1998, responses to this question were grouped into a total of 414 categories (although the average poll in our study reported only 25 categories). Given this information, we grouped the data into 26 categories which also served as the categories into which news stories were classified. Figure 1 shows the categories that we developed for this study. A breakdown of a few of the categories into their specific responses is given in Figure 2.

Like other researchers who have used these data or similar data [5, 1] we had to deal with the problem of inconsistent coding. For example, in some surveys "Medicare" and "Social Security" were considered two separate categories and in others they were reported as one category. This forced us to use a higher level of aggregation than we might have liked. Another major issue is that some of the surveys allowed respondents to give more than one answer. We dealt with this problem by normalizing all of the totals to 100 percent.

4 News data

Our goal in gathering news data was to collect a set of news that would cover the 1990s as thoroughly as possible. We then built and ran a set of classifiers for the same topics mentioned above in processing the poll data.

4.1 Gathering news

In collecting news stories for this task, we used the following guidelines to decide what to include:

- *Heterogeneity.* Using collections from a range of news services should reduce the bias effect of specialized publications. For example, the Wall Street Journal has a heavy concentration of financial reporting that should be counterbalanced with different sources.

AB	Abortion
CO	Cost of living, Inflation
CR	Crime, violence
DI	Dissatisfaction with government
DO†	Don't know, Refused, None, No opinion
DR	Drugs, Drug abuse
EC	Economy (general)
ED	Education, Schools
EN	Environment, Pollution
ET	Ethical, Moral, Family decline, Religious decline
FE	Federal budget deficit, Federal debt, Failure to balance budget
FO	Foreign aid, Focus overseas
FU	Fuel, Oil Prices
GP*	Government, Politics
GU	Guns, gun control
HE	Health care, hospitals, cost of health care
IN	International problems, Foreign affairs
ME	Medicare increases, senior citizens insurance, Social Security
MI	Military, National security, Terrorism, defense issues
OT	Other
PO	Poverty, Hunger, Homelessness, Welfare
RA	Race relations, Racism
RE	Recession, Depression
SP*	Sports
TA	Taxes
TR	Trade deficit, Trade relations, Balance of trade, Foreign trade
UN	Unemployment, jobs, wage issues
WA	Fear of war, Nuclear war

Figure 1: Categories of poll answers, representing a classification respondents' answer for the most important problem facing the nation. A * means the category was used for news only and not to categorize the poll responses. A † means that the category was used in polls but never annotated in the news.

[2] The exact wording of the question has varied somewhat over time [11] but remained the same over the period of our study.

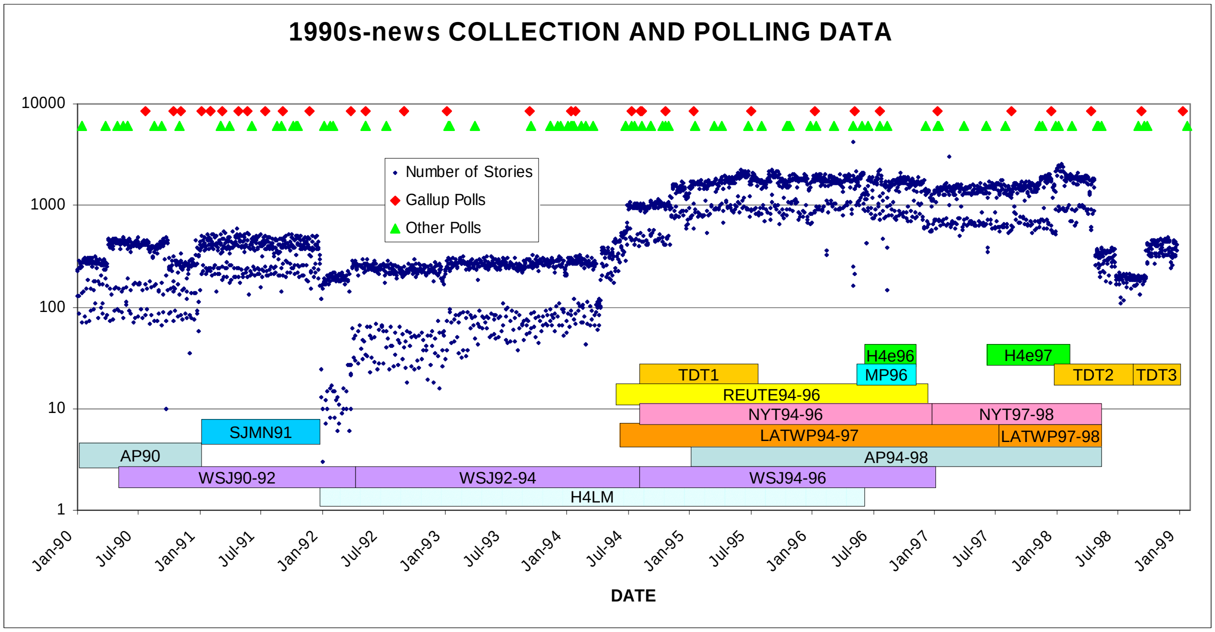

Figure 3: Statistics about poll and news coverage used in this study.

- *Multi-modality.* News articles published in newspapers are only a small part of how people receive news. In fact, more people get their news from television and radio than newspaper. We therefore felt it was important to include transcripts from audio sources as much as possible. Unfortunately, these sources are more difficult to acquire, so they do not represent as large a portion of our final collection as we would have preferred.

- *Coherence.* The collected news needs to discuss the events that are likely to be of interest to the people who answered the poll questions—viz., English speaking US residents. Although some of our sources are from England, they generally discuss events that are of interest to the American public.

With those guidelines in mind, we built the *1990s-news* corpus using all news story collections that were available from the Linguistic Data Consortium[3] that included stories in the 1990s. We excluded any documents marked as commercials in the TDT2, TDT3, MP96, H4e96, and H4e97 collections.

The *1990s-news* corpus includes 2,374,953 stories from 18 collections incorporating 14 sources. Although only half of the sources are newswire (text), they account for about 85% of the total stories. Figure 4 lists the names of the collections and the sources that are included and Figure 5 shows a breakdown of how the stories are distributed across the collections. Figure 3 shows the time period covered by each collection as well as the total number of stories per day throughout the decade covered.

The *1990s-news* corpus is not entirely perfect by our criteria. Coverage is weak in 1992 through 1994 and in much of 1998, though there are sufficient stories on most days: the days with fewer than 100 stories in 1990 and in 1992-1994 are weekends when all sources reduce their reporting. The Reuters col-

[3] http://www.ldc.upenn.edu

DI — Dissatisfaction with government
(bill) clinton scandal
george bush
ronald reagan
bush
campaign finance/campaign money
congress/government
corruption in government
democrats
failure of government
fix government
leadership
miscellaneous government issues
partisan politics
republicans
republicans/gingrich/dole
u.s. senate/house of representatives

GU — Guns, gun control
availability of guns
gun control
gun control/gun laws too strong
gun laws too weak/availability of guns
guns
guns, gun control
guns/gun control
too many guns/gun control

TA — Taxes
high taxes
spending and taxes
taxes

Figure 2: All poll responses that were included for the GU and TA categories, and a sampling of those included in the DI category.

Set	Sources
AP90	Associated Press
AP94-98	Associated Press
H4e96	ABC†, CNN†, CSPAN†, NPR⋄, PRI⋄
H4e97	ABC†, CNN†, CSPAN†, NPR⋄, PRI⋄
H4LM	CNN†, NPR⋄
LATWP94-97	LA Times, Washington Post
LATWP97-98	LA Times, Washington Post
MP96	NPR⋄ (Marketplace)
NYT94-96	NYT
NYT97-98	NYT
REUTE94-96	Reuters
SJMN91	San Jose Mercury News
TDT1	CNN†, Reuters
TDT2	NYT, Associated Press, CNN†, ABC†, PRI⋄, Voice of America⋄
TDT3	NYT, Associated Press, CNN†, ABC†, PRI⋄, Voice of America⋄, NBC†
WSJ90-92	Wall Street Journal
WSJ92-94	Wall Street Journal
WSJ94-96	Wall Street Journal

Figure 4: List of sources included in corpus. Television sources are marked with † and radio sources are marked with ⋄.

lection is from a British source and may not accurately reflect the news selected for American audiences. And the Voice of America broadcasts included in the TDT collections are *by design* targeted for a non-American audience. We felt that acquiring additional data was worth the small problems that may be introduced by those collections.

A particularly time consuming part of the collection gathering was standardizing the collection formats. We needed a standard format so that we could rapidly separate our collection into training and test sets and so that all of the data could comply with the formats required by the text categorization software packages that we used. The formatting also provided us an opportunity to ensure that every document had an identifier that was unique across collections.

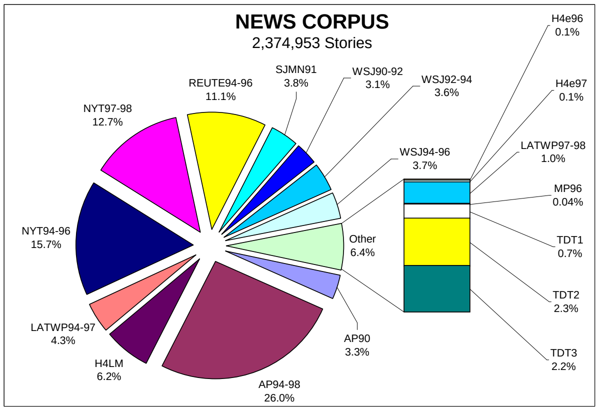

Figure 5: Proportion of news stories by source

First we used custom-built csh and awk scripts to transform the original collections (generally in some form of SGML) into an XML compliant document. This process was very tedious because of the numerous formatting errors, incomplete tags, invalid characters (for the iso-8895-1 character set), repeated documents, non-unique document identifiers, and so on. Without question, the process of cleaning the data consumed the largest portion of the work on this project.

Finally, we built a Java-based XML parser based on Sun Microsystem's implementation [12] of the SAX (Simple API for XML) standard [6]. Because every collection had its own set of tags, we customized the documentHandler class on a collection-by-collection basis, to convert the XML files into the format required by the categorization software. The same process assigned a date to each story, usually derived from the document ID or a date XML tag. The date was used to align news stories and opinion polls.

4.2 Categorization of news

For each of the 26 categories of poll responses, we wish to know which of the news stories discuss events related to those categories. For example, stories discussing the balance of trade are related to the TR (trade) category, stories about the high cost of gasoline fall into FU (fuel oil), and stories about the US government's deficit are grouped into FE (federal debt). Past studies have done this classification by hand [3, 7], severely limiting the amount of news that can be classified. One study devoted 10,000 hours of effort to their classification project [7] and still had a small set of stories categorized. Although we did not keep careful track of our time, we estimate that we used about 250 hours to prepare the data. The categorization itself took less than another 40, almost all of which was the bootstrapping manual classification. Given a set of manually classified topics, the actual categorization of the 2.37 million documents takes approximately four hours.

In this study, we apply automatic text categorization techniques. To bootstrap the process we indexed the SJMN collection from 1991 with InQuery [2]. We then used the description words for each category (see Figure 1) as queries and retrieved a set of stories. We manually judged the relevance of each retrieved story, looking in particular for positive examples for the category. A story could be relevant to multiple topics—in particular, stories in economic categories were difficult to put into a single category.

From that point, we iterated two processes to find sufficient examples for the categories:

1. For a category X, a classifier was trained on all stories marked as being on that topic. The classifier was then run on all collections in our data set. The most highly scoring stories in the entire collection were judged for relevance to the topic, another topic, or to *no* topics at all (represented by the pseudo topic "OT, Other"). We used the BoosTexter system developed at AT&T Labs Research [10, 8], available on-line for research use [9].

2. A person would enter queries to find on-topic stories using his or her knowledge of the topic and run them against various collections of the data set. For example, to find stories on the "HE, health care" topic, a person might enter queries such as "health insurance" or "single payer" or "uninsured", each of which are likely to mentioned in a story discussing health care related

Table 1: Shows how many on-topic stories were identified (manually) for each of the categories used. Note that these are the same categories as in Figure 1, except the unused ones are omitted.

Code	Count	Code	Count	Code	Count
AB	88	FE◇	71	OT	322
CO◇	42	FO	26	PO	59
CR	58	FU◇	28	RA	30
DI	17	GP*	125	RE◇	16
DR	67	GU	43	SP*	217
EC◇	81	HE	52	TA◇	39
ED	71	IN	142	TR◇	37
EN	64	ME	54	UN	36
ET	86	MI	29	WA	25

problems. The top-ranked stories were examined to find additional on-topic stories.

During the process, we noticed that there were classes of stories that were common in the news but that had no relationship to the poll responses. For example, sports stories appear frequently, but are not particularly associated with major national problems (for most people). Rather than lump those stories into the “other” category, we created special categories to attract the stories away from other topics: GP for government and politics stories that do not deal with problems otherwise and SP for sports stories that just report on porting events.

The result of this process was a set of 1619 stories known to be relevant to each topic; a story was relevant to 1.2 topics on average. Table 1 shows the count of stories for each topic.

We then constructed classifiers using those training stories and ran them on the entire 2.37 million stories in the corpus. Figure 6 shows how the automatically assigned topics are distributed in three of the collections included. The point of the graph is to show that the distribution of topics is what might be expected, that it varies from topic to topic, and that it is generally the same for collections in the same time period. For example, there is lots of reporting on international stories (IN) and sports (SP), and a great many stories are classified as “other” (OT).

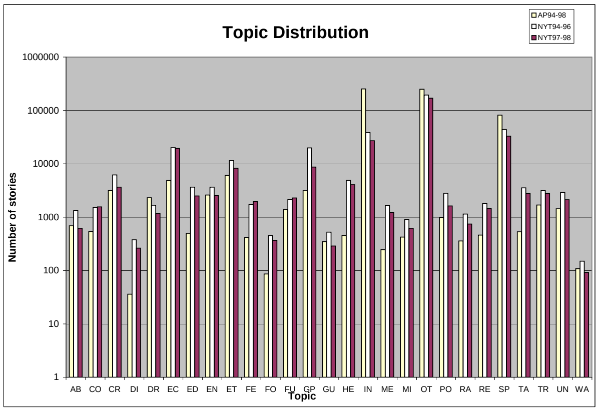

Figure 6: Distribution of topic assignments for three corpora in the data set.

5 Relationships

For each response topic, we now have the percent of people who felt that topic was the most important problem facing the nation on each of about 100 days. We also have, for almost every day from the start of 1990 to the end of 1998, the number of stories that were published on that topic. Because reporting in the *1990s-news* corpus is uneven, we use the percent of stories on that topic per day rather than the raw count. We actually used a seven-day moving average to account for the substantial drop in reporting on weekends. Figure 6 suggests that the distributions are comparable across different subsets of the data; we postulate that as more collections are added to the entire data set, the same trends will hold.

In this section we discuss the regression analysis that we carry out. We first describe the process used to prepare for and carry out the regression, and then we explore several models for relationships between reporting and opinion. We used SPSS version 11 to carry out the analysis.

5.1 Regression variables

There are few poll results and they are fixed in time, so we use them as the starting point for our regression variables. For each topic on the day of a poll, we calculate the following variables:

- P_xx, the percent of poll respondents who felt that *xx* (e.g., HE, IN, GU) was currently the most important problem facing the nation.

- N_xx_1WK, the percent of news stories in the week *before* the polling date that were classified on topic *xx*.

- N_xx_2WK, N_xx_3WK, N_xx_4WK, the same but for the second, third, and fourth weeks before the polling date. Note that the value for the second week is not inclusive of the reporting on the first or third weeks.

5.2 Does news affect polls?

The first question is whether reporting on a particular category has a relationship to the future poll responses. We consider three different ways of approaching the regression:

1. *Within-topic all variables.* We develop a model that uses all four variables (1WK, ..., 4WK) for a given topic, regardless of whether they are needed. This is called "enter" mode in SPSS.

2. *Within-topic stepwise.* We develop the model using all four variables, but they are tried one at a time and only included in the model if they improve its quality. Variables that were added at one point can be removed later if that helps.

3. *Across-topic stepwise.* This is the same as the second option, but we include the reporting numbers from all topics. This allows us to detect, for example, that polling results on gun control are influenced by news reporting on crime (or fuel oil prices).

5.2.1 Within-topic models

Table 2 shows the adjusted R^2 values of the best model for each of the 24 topics, using both of the first two methods. (R^2 is the percentage of variation in the dependent variable predicted by the independent variables. It ranges from 1.0 for perfect prediction to 0.0 for no relationship whatsoever.)

For the within-topic all variables approach, of the 24 categories considered, half showed that the response was dependent on the reporting ($p < 0.05$) and an additional three showed less significant relationships. The R^2 values are rarely very high, which is not surprising since news reporting is unlikely to be the only factor determining opinion.

When the same four reporting variables are used but the model is built using stepwise regression, four additional topics (CO, ET, FE, and TA) show significantly predictive models, meaning that 2/3 of the topics' poll results are predicted by news reporting.

To see how different the models are, we consider three of the topics, selected to be representative: drugs, fuel oil, and gun control. The following table shows the standardized coefficient for each of the variables in the model (i.e., news reporting one through four weeks before the poll). These correspond to the R^2 values listed in the middle column of Table 2.

	DR	FU	GU
1 week	0.160	0.038	0.033
2 weeks	0.074	0.609	-0.080
3 weeks	0.305	-0.654	0.609
4 weeks	0.272	0.499	-0.321

The stepwise regression method only includes a variable in the model if it improves the fit of the model. The following tables lists the final standardized coefficients for each of the models. These correspond to the R^2 values listed in the rightmost column of Table 2.

	DR	FU	GU
1 week	–	–	–
2 weeks	–	0.484	–
3 weeks	0.401	–	0.601
4 weeks	0.367	–	-0.342

In the case of the drugs (DR) model, the terms with low coefficients were dropped out in the stepwise ver-

Topic	All vars	Stepwise
AB	no model	no model
CO	0.038	***0.044**
CR	no model	no model
DI	no model	no model
DR	*0.431	*0.425
EC	*0.168	*0.189
ED	no model	no model
EN	*0.275	*0.280
ET	†0.052	***0.063**
FE	0.018	***0.038**
FO	0.014	no model
FU	*0.269	*0.227
GU	*0.261	*0.272
HE	*0.303	*0.252
IN	0.002	no model
ME	0.016	no model
MI	no model	no model
PO	*0.119	*0.114
RA	*0.258	*0.271
RE	*0.083	*0.084
TA	†0.048	***0.065**
TR	*0.103	*0.108
UN	*0.219	*0.249
WA	*0.081	*0.079

Table 2: Adjusted R^2 values for the best fitting models using news reporting on a topic in the prior four weeks to predict the value of poll response on the same topic (within topic). An asterisk represents significance at $p < 0.05$; a † represents significance at $p < 0.10$. Entries in bold in the second column represent improvements in significance. The value "no model" indicates that SPSS could not find a model with a positive correlation.

sion, but the model is otherwise the same. If true, this model suggests that reporting on drug issues has an impact on future opinion about the important of that topic, but that very recent reporting is not critical.

The change in the fuel oil (FU) model is less obvious. The low coefficient variable is dropped, but then it appears that the reporting in the 3rd and 4th week cancel out, making the 2nd week (the most correlated) the best predictor.

The gun control (GU) model also drops two poorly related variables with low coefficients, but includes the negatively correlated information. This model is suspect because the polling results are almost always low. Despite the press that gun control issues occasionally receive in the US, the poll responses indicating that gun control is the major problem facing the nation never go above 2%. Since the numbers are so small, the variation is small, and it may be possible for a wide range of models to predict it.

5.2.2 Across-topic models

It is very clear that there are dependencies *across* the categories as well as between them. For example, there are obvious relationships between unemployment, hunger, and health care: it is likely that reporting on one topic might impact feelings about the importance of another. For that reason, we ran stepwise regression, trying to predict a particular topic's poll numbers using reporting on *all* topics in the prior four weeks. The R^2 values were often higher in this case, though the models are somewhat suspect because the regression analysis does not properly take the dependencies into account.

We describe two of the models as examples of what can be found. The model that predicts opinion on health care (HE) has an adjusted R^2 of 0.566, one of the higher values we've seen ($p < 0.001$). It includes the following coefficients:

0.319	HE (health care), 2 weeks ago
0.469	ET (ethics), 3 weeks ago
0.229	ME (medicare), 1 week ago
0.220	DI (dissatisfied with government), 3 weeks ago

The first and third seem quite reasonable and strongly related to opinions on health care, though it is unclear how ethics is related. Dissatisfaction with government may be related because it was during the 1990s that the first Clinton administration developed a national health care plan that upset many people. The drug use (DR) response builds a similarly complex model:

0.338	DR (drug use), 3 weeks ago
0.116	TR (trade deficit), 4 weeks ago
0.147	CR (crime), 2 weeks ago
0.161	EC (economics), 3 weeks ago
0.184	CO (cost of living), 4 weeks ago
0.443	MI (military), 4 weeks ago

This model seems to suggest that economic factors (including the cost of living and the trade deficit) play a large part in people's thinking that drug use is a major problem. This is not entirely unreasonable in that a worsening economy seems to be related to increased drug use. We suspect that the military connection is because of the connection between terrorist organizations and illegal drugs, as well as because of the military's involvement in drug policing (the "war on drugs").

5.3 Do polls affect news?

Although prior work strongly suggests that reporting affects public opinion, we felt it would be interesting to explore the reverse correlation. At least one other study found a relationship in this direction [1]. Given the amount of reporting on a particular topic on some day, we looked at the poll responses from one, two, three, or four weeks earlier—that is, a parallel to the other analysis just discussed.

We have news reporting for about 3,000 days but only 102 polls. In order to figure out the response rate for a topic on a particular day, we linearly interpolated the response rates for periods when there were no polls. This assumes a smooth change in public opinion which is clearly incorrect, but it is the best that we can manage with the data available.

We only looked at whether the response rate on a particular topic impacted reporting on the same topic; that is, we did not consider any across-topic effects. We found a few of the variables had an R^2 value over 0.01, though because there were so many data points, the correlation at even such low a level was always statistically significant.

As an example of the sort of model derived, consider one of the handful that did have an adjusted R^2 over 0.01. The drug use (DR) model had a value of 0.297 and the stepwise model included these standardized coefficients:

0.349	polling, 4 weeks ago
0.203	polling, 1 week ago

The coefficients for the unused variables were negative (for 2 weeks ago) or very small (for 3 weeks ago). The accuracy of this model is, of course, suspect for the reasons outlined above. However, it is quite plausible that news reporting is somewhat affected by opinions of what is important, and major issues such as drug use and abuse are likely to show that effect.

Other topics that showed a reasonable correlation were economics (EC) at 0.329, fuel oil (FU) at 0.240, health care (HE) at 0.155, and unemployment (UN) at 0.163.

6 Conclusion

We have shown that automatic text classification methods can be used to greatly expand the volume of data available for a particular kind of social science research. As part of that process, we have constructed *1990s-news*, a mamoth evaluation corpus of over 2.37 million news stories, using resources that are readily available to researchers.

The regression that we have done shows that news reporting on a topic is strongly predictive of the public's opinion of the importance of that topic as a national problem. This result is consistent with past work in social science, suggesting that our corpus construction process was done correctly.

We are pleased by the work to date, but feel that substantially more work could be done to improve the quality of the classifiers. We suspect that some of the unusual results in the regression models is because of classification errors that bring somewhat similar topics together incorrectly. We are currently modifying the classification approach to allow a story to fall into multiple topics, requiring more careful work on threshold selection (to better identify stories that are on none of our topics).

There is reason to believe that the approach we are using for regression analysis will produce upwardly biased coefficients. Specifically, the problem is that the responses to the question are not independent: an increase in the percentage indicating that health care is a problem must cause a corresponding decrease in the other categories. To help with this problem, we are intending to explore two-stage estimators [1].

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by a grant from the Historical Research Foundation and Dartmouth College. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] Roy L. Behr and Shanto Iyengar. Television news, real-world cues, and changes in the public agenda. *Public Opinion Quarterly*, 49(1):38–57, Spring 1985.

[2] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of DEXA-92, 3rd International Conference on Datab ase and Expert Systems Applications*, pages 78–83, 1992.

[3] Donald L. Jordan. Newspaper effects on policy preferences. *Public Opinion Quarterly*, 57(2):191–204, Summer 1993.

[4] Donald L. Jordan and Benjamin I. Page. Shaping foreign policy opinions: The role of tv news. *The Journal of Conflict Resolution*, 36(2):227–241, June 1992.

[5] Maxwell McCombs and Jian-Hua Zhu. Capacity, diversity, and volatility of the public agenda: Trends from 1954 to 1994. *Public Opinion Quarterly*, 59(4):495–525, Winter 1995.

[6] David Megginson. Simple API for XML, 2000. http://www.saxproject.org/.

[7] Benjamin I. Page, Robert Y. Shapiro, and Glenn R. Dempsey. What moves public opinion? *The American Political Science Review*, 81(1):23–44, March 1987.

[8] R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 80–91, 1998.

[9] Robert Schapire. Boostexter, 2000. http://www.research.att.com/ schapire/BoosTexter/.

[10] Robert E. Shapire and Yoram Singer. BoosTexter: A boosting based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[11] Tom W. Smith. The polls: America's most important problems part i: National and international. *Public Opinion Quarterly*, 49(2):264–274, Summer 1985.

[12] Inc. Sun Microsystems. Java API for XML processing (jaxp), 2000. http://java.sun.com/xml/jaxp/index.html.