2010

From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

Brendan O'Connor
Carnegie Mellon University

Ramnath Balasubramanyan
Carnegie Mellon University

Bryan R. Routledge
Carnegie Mellon University, routledge@cmu.edu

Noah A. Smith
Carnegie Mellon University, nasmith@cs.cmu.edu

Published In
Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

Brendan O'Connor[†] **Ramnath Balasubramanyan**[†] **Bryan R. Routledge**[§] **Noah A. Smith**[†]
brenocon@cs.cmu.edu rbalasub@cs.cmu.edu routledge@cmu.edu nasmith@cs.cmu.edu

[†]School of Computer Science
Carnegie Mellon University

[§]Tepper School of Business
Carnegie Mellon University

Abstract

We connect measures of public opinion measured from polls with sentiment measured from text. We analyze several surveys on consumer confidence and political opinion over the 2008 to 2009 period, and find they correlate to sentiment word frequencies in contemporaneous Twitter messages. While our results vary across datasets, in several cases the correlations are as high as 80%, and capture important large-scale trends. The results highlight the potential of text streams as a substitute and supplement for traditional polling.

Introduction

If we want to know, say, the extent to which the U.S. population likes or dislikes Barack Obama, an obvious thing to do is to ask a random sample of people (i.e., poll). Survey and polling methodology, extensively developed through the 20th century (Krosnick, Judd, and Wittenbrink 2005), gives numerous tools and techniques to accomplish representative public opinion measurement.

With the dramatic rise of text-based social media, millions of people broadcast their thoughts and opinions on a great variety of topics. Can we analyze publicly available data to infer population attitudes in the same manner that public opinion pollsters query a population? If so, then mining public opinion from freely available text content could be a faster and less expensive alternative to traditional polls. (A standard telephone poll of one thousand respondents easily costs tens of thousands of dollars to run.) Such analysis would also permit us to consider a greater variety of polling questions, limited only by the scope of topics and opinions people broadcast. Extracting the public opinion from social media text provides a challenging and rich context to explore computational models of natural language, motivating new research in computational linguistics.

In this paper, we connect measures of public opinion derived from polls with sentiment measured from analysis of text from the popular microblogging site Twitter. We explicitly link measurement of textual sentiment in microblog messages through time, comparing to contemporaneous polling data. In this preliminary work, summary statistics derived from extremely simple text analysis techniques are demonstrated to correlate with polling data on consumer confidence and political opinion, and can also predict future movements in the polls. We find that temporal smoothing is a critically important issue to support a successful model.

Data

We begin by discussing the data used in this study: Twitter for the text data, and public opinion surveys from multiple polling organizations.

Twitter Corpus

Twitter is a popular microblogging service in which users post messages that are very short: less than 140 characters, averaging 11 words per message. It is convenient for research because there are a very large number of messages, many of which are publicly available, and obtaining them is technically simple compared to scraping blogs from the web.

We use 1 billion Twitter messages posted over the years 2008 and 2009, collected by querying the Twitter API,[1] as well as archiving the "Gardenhose" real-time stream. This comprises a roughly uniform sample of public messages, in the range of 100,000 to 7 million messages per day. (The primary source of variation is growth of Twitter itself; its message volume increased by a factor of 50 over this two-year time period.)

Most Twitter users appear to live in the U.S., but we made no systematic attempt to identify user locations or even message language, though our analysis technique should largely ignore non-English messages.

There probably exist many further issues with this text sample; for example, the demographics and communication habits of the Twitter user population probably changed over this time period, which should be adjusted for given our desire to measure attitudes in the general population. There are clear opportunities for better preprocessing and stratified sampling to exploit these data.

[1]This scraping effort was conducted by Brendan Meeder.

Public Opinion Polls

We consider several measures of consumer confidence and political opinion, all obtained from telephone surveys to participants selected through random-digit dialing, a standard technique in traditional polling (Chang and Krosnick 2003).

Consumer confidence refers to how optimistic the public feels, collectively, about the health of the economy and their personal finances. It is thought that high consumer confidence leads to more consumer spending; this line of argument is often cited in the popular media and by policymakers (Greenspan 2002), and further relationships with economic activity have been studied (Ludvigson 2004; Wilcox 2007). Knowing the public's consumer confidence is of great utility for economic policy making as well as business planning.

Two well-known surveys that measure U.S. consumer confidence are the Consumer Confidence Index from the Consumer Board, and the Index of Consumer Sentiment (ICS) from the Reuters/University of Michigan Surveys of Consumers.[2] We use the latter, as it is more extensively studied in economics, having been conducted since the 1950s. The ICS is derived from answers to five questions administered monthly in telephone interviews with a nationally representative sample of several hundred people; responses are combined into the index score. Two of the questions, for example, are:

> "We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?"
>
> "Now turning to business conditions in the country as a whole—do you think that during the next twelve months we'll have good times financially, or bad times, or what?"

We also use another poll, the Gallup Organization's "Economic Confidence" index,[3] which is derived from answers to two questions that ask interviewees to to rate the overall economic health of the country. This only addresses a subset of the issues that are incorporated into the ICS. We are interested in it because, unlike the ICS, it is administered daily (reported as three-day rolling averages). Frequent polling data are more convenient for our comparison purpose, since we have fine-grained, daily Twitter data, but only over a two-year period. Both datasets are shown in Figure 1.

For **political opinion**, we use two sets of polls. The first is Gallup's daily tracking poll for the presidential job approval rating for Barack Obama over the course of 2009, which is reported as 3-day rolling averages.[4] These data are shown in Figure 2.

The second is a set of tracking polls during the 2008 U.S. presidential election cycle, asking potential voters whether they would vote for Barack Obama or John McCain. Many different organizations administered them throughout 2008; we use a compilation provided by Pollster.com, consisting of 491 data points from 46 different polls.[5] The data are shown in Figure 3.

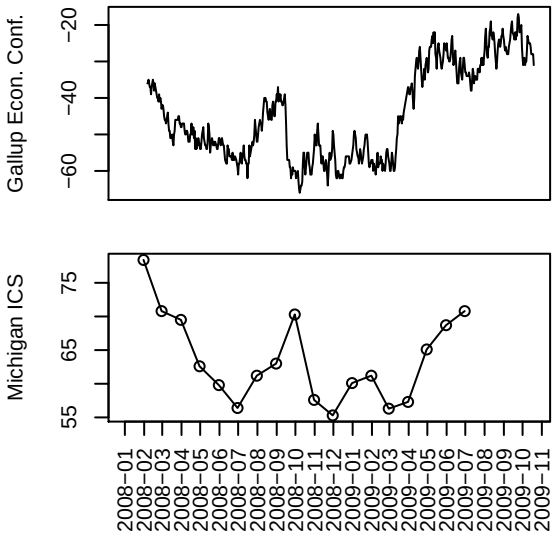

Figure 1: Monthly Michigan ICS and daily Gallup consumer confidence poll.

Text Analysis

From text, we are interested in assessing the population's aggregate opinion on a topic. Immediately, the task can be broken down into two subproblems:

1. Message retrieval: identify messages relating to the topic.
2. Opinion estimation: determine whether these messages express positive or negative opinions or news about the topic.

If there is enough training data, this could be formulated as a topic-sentiment model (Mei et al. 2007), in which the topics and sentiment of documents are jointly inferred. Our dataset, however, is asymmetric, with millions of text messages per day (and millions of distinct vocabulary items) but only a few hundred polling data points in each problem. It is a challenging setting to estimate a useful model over the vocabulary and messages. The signal-to-noise ratio is typical of information retrieval problems: we are only interested in information contained in a small fraction of all messages.

We therefore opt to use a transparent, deterministic approach based on prior linguistic knowledge, counting instances of positive-sentiment and negative-sentiment words in the context of a topic keyword.

[2]Downloaded from `http://www.sca.isr.umich.edu/`.

[3]Downloaded from `http://www.gallup.com/poll/122840/gallup-daily-economic-indexes.aspx`.

[4]Downloaded from `http://www.gallup.com/poll/113980/Gallup-Daily-Obama-Job-Approval.aspx`.

[5]Downloaded from `http://www.pollster.com/polls/us/08-us-pres-ge-mvo.php`

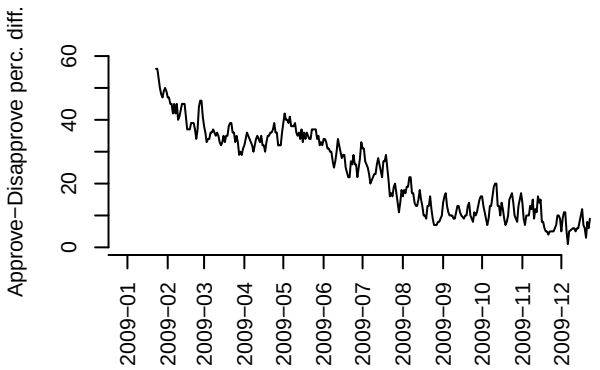

Figure 2: 2009 presidential job approval (Barack Obama).

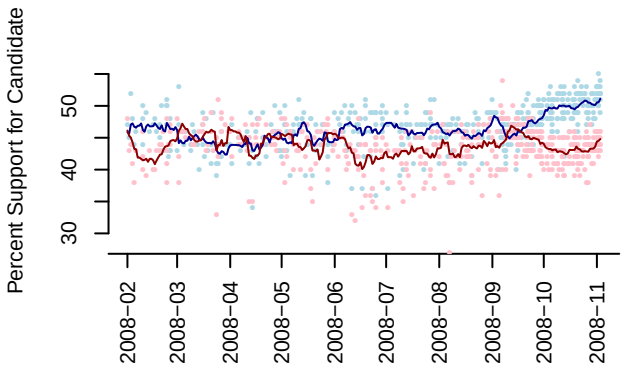

Figure 3: 2008 presidential elections, Obama vs. McCain (blue and red). Each poll provides separate Obama and McCain percentages (one blue and one red point); lines are 7-day rolling averages.

Message Retrieval

We only use messages containing a topic keyword, manually specified for each poll:

- For consumer confidence, we use *economy*, *job*, and *jobs*.
- For presidential approval, we use *obama*.
- For elections, we use *obama* and *mccain*.

Each topic subset contained around 0.1–0.5% of all messages on a given day, though with occasional spikes, as seen in Figure 4. These appear to be driven by news events. All terms have a weekly cyclical structure, occurring more frequently on weekdays, especially in the middle of the week, compared to weekends. (In the figure, this is most apparent for the term *job* since it has fewer spikes.) Nonetheless, these fractions are small. In the earliest and smallest part of our dataset, the topic samples sometimes come out just several hundred messages per day; but by late 2008, there are thousands of messages per day for most datasets.

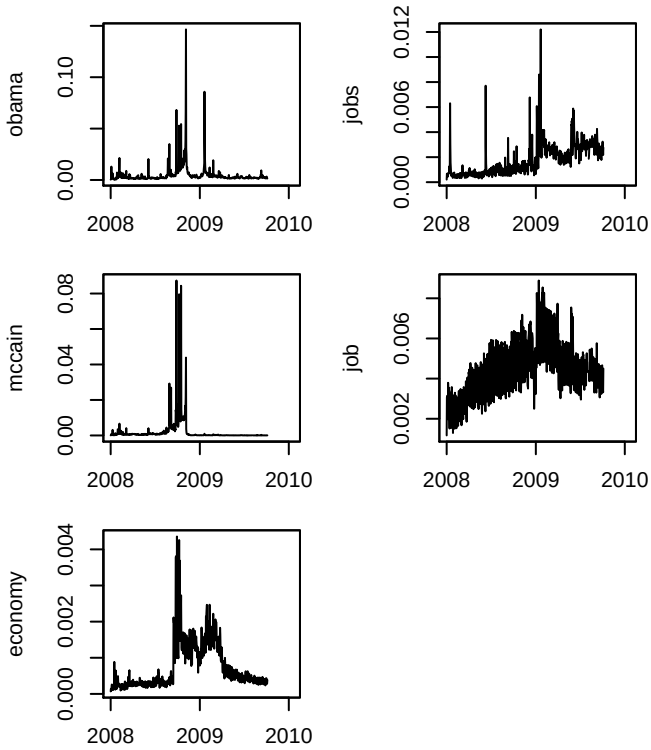

Figure 4: Fraction of Twitter messages containing various topic keywords, per day.

Opinion Estimation

We derive day-to-day sentiment scores by counting positive and negative messages. Positive and negative words are defined by the subjectivity lexicon from OpinionFinder, a word list containing about 1,600 and 1,200 words marked as positive and negative, respectively (Wilson, Wiebe, and Hoffmann 2005).[6] We do not use the lexicon's distinctions between weak and strong words.

A message is defined as positive if it contains any positive word, and negative if it contains any negative word. (This allows for messages to be both positive and negative.) This gives similar results as simply counting positive and negative words on a given day, since Twitter messages are so short (about 11 words).

We define the sentiment score x_t on day t as the ratio of positive versus negative messages on the topic, counting from that day's messages:

$$x_t = \frac{\text{count}_t(\text{pos. word} \wedge \text{topic word})}{\text{count}_t(\text{neg. word} \wedge \text{topic word})} = \frac{p(\text{pos. word} \mid \text{topic word}, t)}{p(\text{neg. word} \mid \text{topic word}, t)} \quad (1)$$

where the likelihoods are estimated as relative frequencies.

We performed casual inspection of the detected messages and found many examples of falsely detected sentiment. For example, the lexicon has the noun *will* as a weak positive word, but since we do not use a part-of-speech tagger, this

[6] Available at `http://www.cs.pitt.edu/mpqa`.

causes thousands of false positives when it matches the verb sense of *will*.[7] Furthermore, recall is certainly very low, since the lexicon is designed for well-written standard English, but many messages on Twitter are written in an informal social media dialect of English, with different and alternately spelled words, and emoticons as potentially useful signals. Creating a more comprehensive lexicon with distributional similarity techniques could improve the system; Velikovich et al. (2010) find that such a web-derived lexicon substantially improves a lexicon-based sentiment classifier.

Comparison to Related Work

The sentiment analysis literature often focuses on analyzing individual documents, or portions thereof (for a review, see Pang and Lee, 2008). Our problem is related to work on sentiment information retrieval, such as the TREC Blog Track competitions that have challenged systems to find and classify blog posts containing opinions on a given topic (Ounis, MacDonald, and Soboroff 2008).

The sentiment feature we consider, presence or absence of sentiment words in a message, is one of the most basic ones used in the literature. If we view this system in the traditional light—as subjectivity and polarity detection for individual messages—it makes many errors, like all natural language processing systems. However, we are only interested in *aggregate* sentiment. A high error rate merely implies the sentiment detector is a noisy measurement instrument. With a fairly large number of measurements, these errors will cancel out relative to the quantity we are interested in estimating, aggregate public opinion.[8] Furthermore, as Hopkins and King (2010) demonstrate, it can actually be inaccurate to naïvely use standard text analysis techniques, which are usually designed to optimize per-document classification accuracy, when the goal is to assess aggregate population proportions.

Several prior studies have estimated and made use of aggregated text sentiment. The informal study by Lindsay (2008) focuses on lexical induction in building a sentiment classifier for a proprietary dataset of Facebook wall posts (a web conversation/microblog medium broadly similar to Twitter), and demonstrates correlations to several polls conducted during part of the 2008 presidential election. We are unaware of other research validating text analysis against traditional opinion polls, though a number of companies offer text sentiment analysis basically for this purpose (e.g., Nielsen Buzzmetrics). There are at least several other studies that use time series of either aggregate text sentiment or good vs. bad news, including analyzing stock behavior based on text from blogs (Gilbert and Karahalios 2010), news articles (Lavrenko et al. 2000; Koppel and Shtrimberg 2004) and investor message boards (Antweiler and Frank 2004; Das and Chen 2007). Dodds and Danforth (2009) use an emotion word counting technique for purely exploratory analysis of several corpora.

[7] We tried manually removing this and several other frequently mismatching words, but it had little effect.

[8] There is an issue if errors correlate with variables relevant to public opinion; for example, if certain demographics speak in dialects that are harder to analyze.

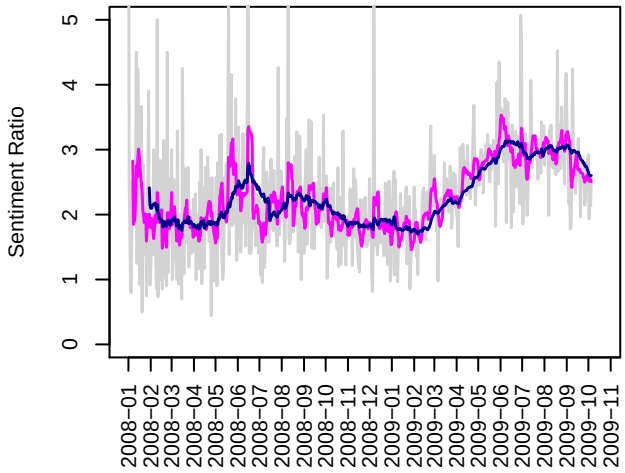

Figure 5: Moving average MA_t of sentiment ratio for *jobs*, under different windows $k \in \{1, 7, 30\}$: no smoothing (gray), past week (magenta), and past month (blue). The unsmoothed version spikes as high as 10, omitted for space.

Moving Average Aggregate Sentiment

Day-to-day, the sentiment ratio is volatile, much more than most polls.[9] Just like in the topic volume plots (Figure 4), the sentiment ratio rapidly rises and falls each day. In order to derive a more consistent signal, and following the same methodology used in public opinion polling, we *smooth* the sentiment ratio with one of the simplest possible temporal smoothing techniques, a moving average over a window of the past k days:

$$MA_t = \frac{1}{k}\left(x_{t-k+1} + x_{t-k+2} + ... + x_t\right)$$

Smoothing is a critical issue. It causes the sentiment ratio to respond more slowly to recent changes, thus forcing consistent behavior to appear over longer periods of time. Too much smoothing, of course, makes it impossible to see fine-grained changes to aggregate sentiment. See Figure 5 for an illustration of different smoothing windows for the *jobs* topic.

Correlation Analysis: Is text sentiment a leading indicator of polls?

Figure 6 shows the *jobs* sentiment ratio compared to the two different measures of consumer confidence, Gallup Daily and Michigan ICS. It is apparent that the sentiment ratio captures the broad trends in the survey data. With 15-day smoothing, it is reasonably correlated with Gallup at $r = 73.1\%$. The most glaring difference is a region of high positive sentiment in May-June 2008. But otherwise, the sentiment ratio seems to pick up on the downward slide of consumer confidence through 2008, and the rebound in February/March of 2009.

[9] That the reported poll results are less volatile does not imply that they are more accurate reflections of true population opinion than the text.

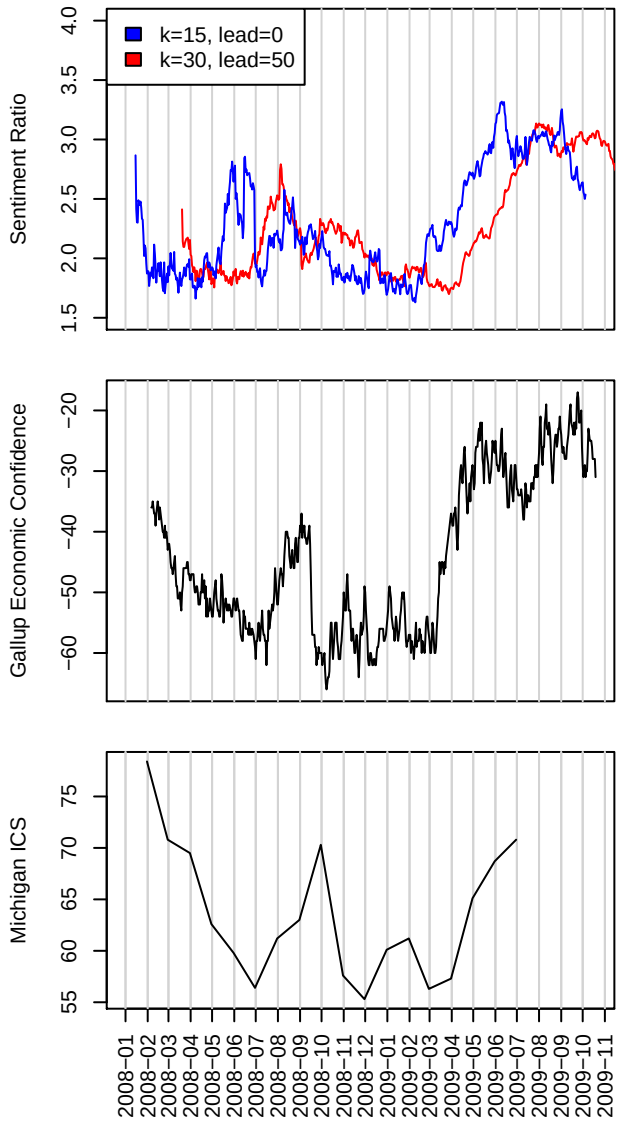

Figure 6: Sentiment ratio and consumer confidence surveys. Sentiment information captures broad trends in the survey data.

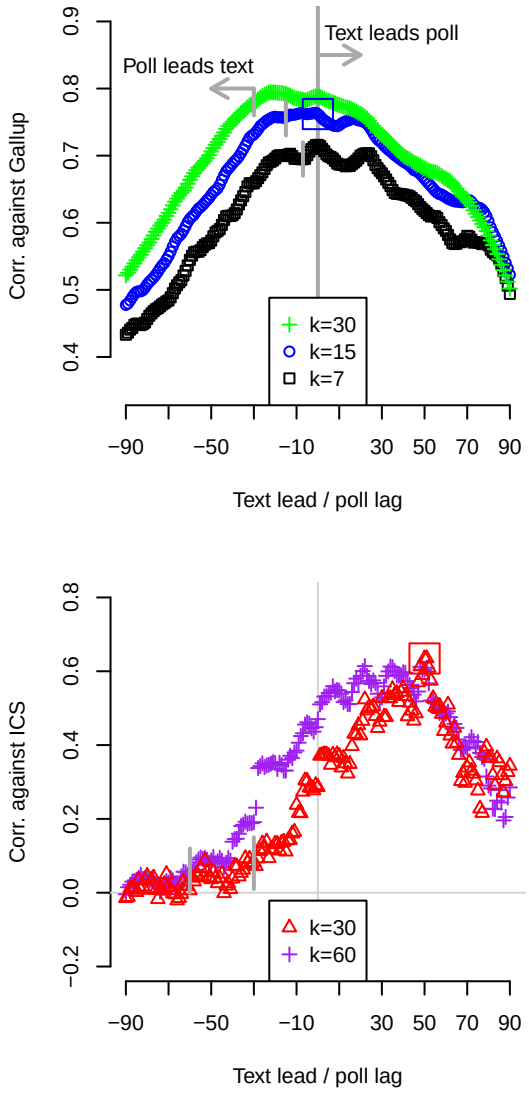

Figure 7: Cross-correlation plots: sensitivity to lead and lag for different smoothing windows. $L > 0$ means the text window completely precedes the poll, and $L < -k$ means the poll precedes the text. (The window straddles the poll for $L < -k < 0$.) The $L = -k$ positions are marked on each curve. The two parameter settings shown in Figure 6 are highlighted with boxes.

When consumer confidence changes, can this first be seen in the text sentiment measure, or in polls? If text sentiment responds faster to news events, a sentiment measure may be useful for economic researchers and policymakers. We can test this by looking at leading versions of text sentiment.

First note that the text-poll correlation reported above is the goodness-of-fit metric for a one variable linear least-squares model:

$$y_t = b + a \sum_{T=t-k+1}^{t} x_T + \epsilon_t$$

for poll outcomes y_t, daily sentiment ratios x_t, Gaussian noise ϵ_t, and a fixed hyperparameter k. A poll outcome is compared to the k-day text sentiment window that ends on the same day as the poll.

We introduce a lag hyperparameter L into the model, so the poll is compared against the text window ending L days before the poll outcome.

$$y_{t+L} = b + a \sum_{T=t-k+1}^{t} x_T + \epsilon_t$$

Graphically, this is equivalent to taking one of the text sentiment lines on Figure 6 and shifting it to the right by L days, then examining the correlation against the consumer confidence polls below.

Polls are typically administered over an interval. The ICS is reported once per month (at the end of the month), and Gallup is reported for 3-day windows. We always consider the last day of the poll's window to be the poll date, which is the earliest possible day that the information could actually be used. Therefore, we would expect both daily measures, Gallup and text sentiment, to always lead ICS, since it measures phenomena occurring over the previous month.

The sensitivity of text-poll correlation to smoothing window and lag parameters (k, L) is shown in Figure 7. The regions corresponding to text preceding or following the poll are marked. Correlation is higher for text leading the poll and not the other way around, so text seems to be a leading indicator. Gallup correlations fall off faster for poll-leads-text than text-leads-poll, and the ICS has similar properties.

If text and polls moved at random relative to each other, these cross-correlation curves would stay close to 0. The fact they have peaks at all strongly suggests that the text sentiment measure captures information related to the polls.

Also note that more smoothing increases the correlation: for Gallup, 7-, 15-, and 30-day windows peak at $r = 71.6\%$, 76.3%, and 79.4% respectively. The 7-day and 15-day windows have two local peaks for correlation, corresponding to shifts that give alternate alignments of two different humps against the Gallup data, but the better-correlating 30-day window smooths over these entirely. Furthermore, for the ICS, a 60-day window often achieves higher correlation than the 30-day window. These facts imply that the text sentiment information is volatile, and if polls are believed to be a gold standard, then it is best used to detect long-term trends.

It is also interesting to consider ICS a gold standard and compare correlations with Gallup and text sentiment. ICS and Gallup are correlated (best correlation is $r = 86.4\%$ if Gallup is given its own smoothing and alignment at $k = 30, L = 20$), which supports the hypothesis that they are measuring similar things, and that Gallup is a leading indicator for ICS. Fixed to 30-day smoothing, the sentiment ratio only achieves $r = 63.5\%$ under optimal lead $L = 50$. So it is a weaker indicator than Gallup.

Finally, we also experimented with sentiment ratios for the terms *job* and *economy*, which both correlate very poorly with the Gallup poll: 10% and 7% respectively (with the default $k = 15, L = 0$).[10]

This is a cautionary note on the common practice of stemming words, which in information retrieval can have mixed effects on performance (Manning, Raghavan, and Schütze 2008, ch. 2). Here, stemming would have conflated *job* and *jobs*, severely degrading results.

Forecasting Analysis

As a further validation, we can evaluate the model in a rolling forecast setting, by testing how well the text-based model can predict future values of the poll. For a lag L, and a target forecast date $t + L$, we train the model only on historical data through day $t - 1$, then predict using the window ending on day t. The lag parameter L is how many days in the future the forecasts are for. We repeat this model fit and prediction procedure for most days. (We cannot forecast early days in the data, since $L + k$ initial days are necessary to cover the start of the text sentiment window, plus at least several days for training.)

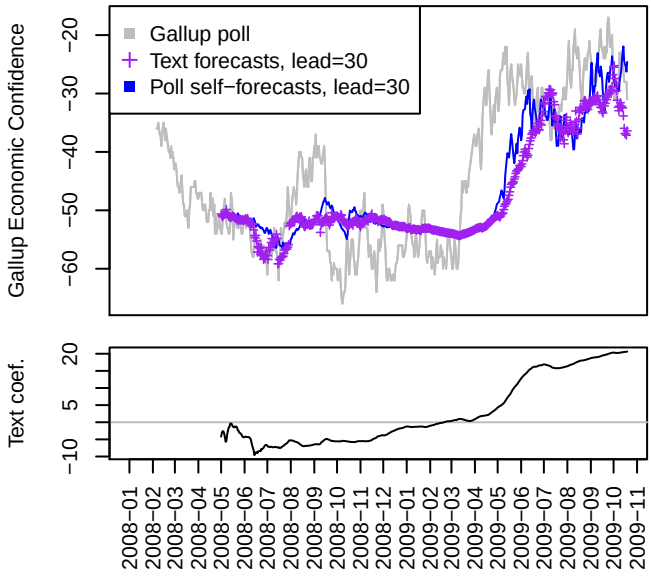

Figure 8: Rolling text-based forecasts (above), and the text sentiment (MA_t) coefficients a for each of the text forecasting models over time (below).

Results are shown in Figure 8. Forecasts for one month in

[10] We inspected some of the matching messages to try to understand this result, but since the sentiment detector is very noisy at the message level, it was difficult to understand what was happening.

the future (that is, using past text from 44 through 30 days before the target date) achieve 77.5% correlation. This is slightly worse than a baseline to predict the poll from its lagged self ($y_{t+L} \approx b_0 + b_1 y_t$), which has $r = 80.4\%$. Adding the sentiment score to historical poll information as a bivariate model ($y_{t+L} \approx b_0 + b_1 y_t + aMA_{t..t-k+1}$), yields a very small improvement ($r = 81.0\%$).

Inspecting the rolling forecasts and text model coefficient a is revealing. In 2008 and early 2009, text sentiment is a poor predictor of consumer confidence; for example, it fails to reflect a hump in the polls in August and September 2008. The model learns a coefficient near zero (even negative), and makes predictions similar to the poll's self-predictions, which is possible since the poll's most recent values are absorbed into the bias term of the text-only model. However, starting in mid-2009, text sentiment becomes a much better predictor, as it captures the general rise in consumer confidence starting then (see Figure 6). This suggests qualitatively different phenomena are being captured by the text sentiment measure at different times. From the perspective of time series modeling, future work should investigate techniques for deciding the importance of different historical signals and time periods, such as vector autoregressions (e.g. Hamilton 1994).

It is possible that the effectiveness of text changes over this time period for reasons described earlier: Twitter itself changed substantially over this time period. In 2008, the site had far fewer users who were probably less representative of the general population, and were using the site differently than users would later.

Obama 2009 Job Approval and 2008 Elections

We analyze the sentiment ratio for *obama* and compared it to two series of polls, presidential job approval in 2009, and presidential election polls in 2008, as seen in Figure 9. The job approval poll is the most straightforward, being a steady decline since the start of the Obama presidency, perhaps with some stabilization in September or so. The sentiment ratio also generally declines during this period, with $r = 72.5\%$ for $k = 15$.

However, in 2008 the sentiment ratio does not substantially correlate to the election polls ($r = -8\%$); we compare to the percent of support for Obama, averaged over a 7-day window of tracking polls: the same information displayed in Figure 3). Lindsay (2008) found that his daily sentiment score was a leading indicator to one particular tracking poll (Rasmussen) over a 100-day period from June-October 2008. Our measure also roughly correlates to the same data, though less strongly ($r = 44\%$ vs. $r = 57\%$) and only at different lag parameters.

The elections setting may be structurally more complex than presidential job approval. In many of the tracking polls, people can choose to answer any *Obama*, *McCain*, *undecided*, *not planning to vote*, and third-party candidates. Furthermore, the name of every candidate has its own sentiment ratio scores in the data. We might expect the sentiment for *mccain* to be vary inversely with *obama*, but they in fact slightly correlate. It is also unclear how they should interact as part of a model of voter preferences.

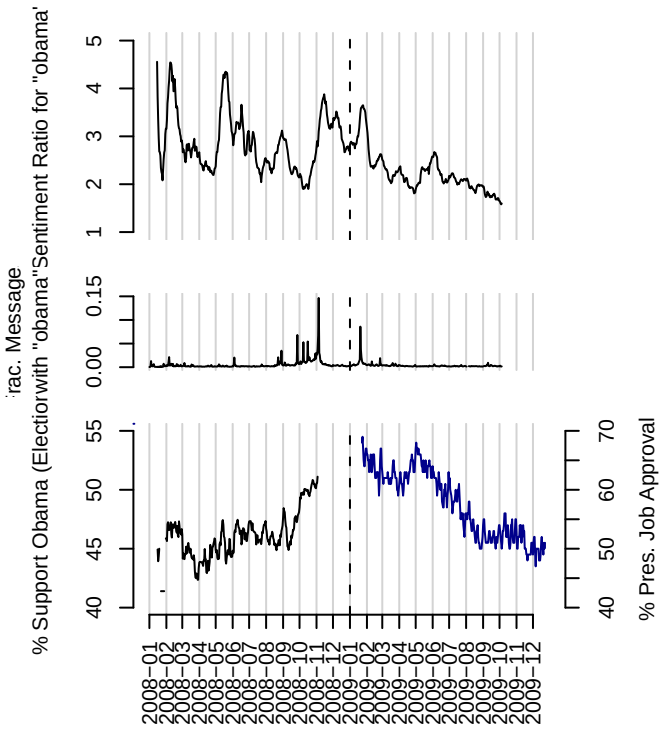

Figure 9: The sentiment ratio for *obama* (15-day window), and fraction of all Twitter messages containing *obama* (day-by-day, no smoothing), compared to election polls (2008) and job approval polls (2009).

We also found that the topic frequencies correlate with polls much more than the sentiment scores. First note that the message volume for *obama*, shown in Figure 9, has the usual daily spikes like other words on Twitter shown in Figure 4. Some of these spikes are very dramatic; for example, on November 5th, nearly 15% of all Twitter messages (in our sample) mentioned the word *obama*.

Furthermore, the *obama* message volume substantially correlates to the poll numbers. Even the raw volume has a 52% correlation to the polls, and the 15-day window version is up to $r = 79\%$. Simple attention seems to be associated with popularity, at least for Obama. But the converse is not true for *mccain*; this word's 15-day message volume *also* correlates to higher Obama ratings in the polls ($r = 74\%$). A simple explanation may be that frequencies of either term *mccain* or *obama* are general indicators of elections news and events, and most 2008 elections news and events were favorable toward or good for Obama. Certainly, topic frequency may not have a straightforward relationship to public opinion in a more general text-driven methodology for public opinion measurement, but given the marked effects it has in these data, it is worthy of further exploration.

Conclusion

In the paper we find that a relatively simple sentiment detector based on Twitter data replicates consumer confidence and presidential job approval polls. While the results do not come without caution, it is encouraging that expensive and

time-intensive polling can be supplemented or supplanted with the simple-to-gather text data that is generated from online social networking. The results suggest that more advanced NLP techniques to improve opinion estimation may be very useful.

The textual analysis could be substantially improved. Besides the clear need for a more well-suited lexicon, the modes of communication should be considered. When messages are retweets (forwarded messages), should they be counted? What about news headlines? Note that Twitter is rapidly changing, and the experiments on recent (2009) data performed best, which suggests that it is evolving in a direction compatible with our approach, which uses no Twitter-specific features at all.

In this work, we treat polls as a gold standard. Of course, they are noisy indicators of the truth — as is evident in Figure 3 — just like extracted textual signals. Future work should seek to understand how these different signals reflect public opinion either as a hidden variable, or as measured from more reliable sources like face-to-face interviews.

Many techniques from traditional survey methodology can also be used again for automatic opinion measurement. For example, polls routinely use stratified sampling and weighted designs to ask questions of a representative sample of the population. Given that many social media sites include user demographic information, such a design is a sensible next step.

Eventually, we see this research progressing to align with the more general goal of query-driven sentiment analysis where one can ask more varied questions of what people are thinking based on text they are already writing. Modeling traditional survey data is a useful application of sentiment analysis. But it is also a stepping stone toward larger and more sophisticated applications.

Acknowledgments

This work is supported by the Center for Applied Research in Technology at the Tepper School of Business, and the Berkman Faculty Development Fund at Carnegie Mellon University. We would like to thank the reviewers for helpful suggestions, Charles Franklin for advice in interpreting election polling data, and Brendan Meeder for contribution of the Twitter scrape.

References

Antweiler, W., and Frank, M. Z. 2004. Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance* 59(3):1259–1294.

Chang, L. C., and Krosnick, J. A. 2003. National surveys via RDD telephone interviewing vs. the internet: Comparing sample representativeness and response quality. Unpublished manuscript.

Das, S. R., and Chen, M. Y. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53(9):1375–1388.

Dodds, P. S., and Danforth, C. M. 2009. Measuring the happiness of Large-Scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies* 116.

Gilbert, E., and Karahalios, K. 2010. Widespread worry and the stock market. In *Proceedings of the International Conference on Weblogs and Social Media*.

Greenspan, A. 2002. Remarks at the Bay Area council conference, San Francisco, California. `http://www.federalreserve.gov/boarddocs/speeches/2002/20020111/default.htm`.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton University Press.

Hopkins, D., and King, G. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1):229–247.

Koppel, M., and Shtrimberg, I. 2004. Good news or bad news? Let the market decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Krosnick, J. A.; Judd, C. M.; and Wittenbrink, B. 2005. The measurement of attitudes. *The Handbook of Attitudes* 2176.

Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; and Allan, J. 2000. Mining of concurrent text and time series. In *Proceedings of the 6th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*.

Lindsay, R. 2008. Predicting polls with Lexicon. `http://languagewrong.tumblr.com/post/55722687/predicting-polls-with-lexicon`.

Ludvigson, S. C. 2004. Consumer confidence and consumer spending. *The Journal of Economic Perspectives* 18(2):29–50.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press, 1st edition.

Mei, Q.; Ling, X.; Wondra, M.; Su, H.; and Zhai, C. X. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International conference on World Wide Web*.

Ounis, I.; MacDonald, C.; and Soboroff, I. 2008. On the TREC blog track. In *Proceedings of the International Conference on Weblogs and Social Media*.

Pang, B., and Lee, L. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Velikovich, L.; Blair-Goldensohn, S.; Hannan, K.; and McDonald, R. 2010. The viability of web-dervied polarity lexicons. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Wilcox, J. 2007. Forecasting components of consumption with components of consumer sentiment. *Business Economics* 42(4):2232.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*.