Extending Opinion Polls through the Combination of Analytical and Simulation Models

Bruce Skarin, Ian Yohai, Robert McCormack
Aptima, Inc.
12 Gill St. Suite 1400, Woburn, MA 01801
781.935.3966
bskarin@aptima.com, iyohai@aptima.com, rmccormack@aptima.com

Abstract

This paper describes a method for extending the accuracy of opinion polls by combining a simulation model with text analysis. Opinion polls are an important tool for gauging how societies interpret issues ranging from elections and policy to uprisings and regime change. While such polls produce detailed data, they are also costly and take time to field. The "Arab Spring" uprisings highlight the challenges this lag creates when governments collapse more quickly than the time required for polling and analysis. In relation to this rapid change is the rise of social networking and the instantaneous information it provides. Yet research indicates that opinions expressed in social media are often not representative of societies as a whole. To integrate these two very different data sources, we use events extracted from media to perturb a simulation of representative agents that were initialized using a prior poll. Agents update opinions using equations developed in a system dynamics model of social identity theory's bounded confidence. We then evaluate the model's performance using longitudinal opinion studies. Preliminary results suggest that the integrated results offer improvements over the sources used in isolation, which may help leaders do better at anticipating important societal changes.

Introduction: Limitations of polls and social media analysis

For over sixty years traditional surveys have been the gold standard in opinion research. These opinion polls represent a detailed description of how a population perceives the state of the world and thus provide insights into how the public may behave given certain situations. From the purchases of specific goods, to votes in an election, to activism for policy or regime change, opinion polls have played a key role in anticipating the outcomes of such events. These polls, however, can only capture a static snapshot of a population living in a world that is changing ever faster. As soon as a survey is completed, the accuracy to which it represents the current state of the population quickly begins to decay. Events, such as competitor product releases, debates, protests, attacks, or even natural disasters, can drastically affect opinions. As such, surveys conducted just a few months past may often be an unreliable representation of the affected population. This is especially true in regions facing conflict or other situations of high volatility.

To help fill these gaps, a significant amount of attention is being given to the growing popularity of social media and twenty-four hour news coverage. For instance Asur and Humberman (2010) was able to "use the chatter from Twitter.com to forecast box-office revenues for movies" and "show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors." Other studies have shown noteworthy results in estimating a variety of other population states including flu prevalence (Corey et al., 2009) and the personality of an individual (Golbeck et al., 2011).

Yet a year-long study by Mitchell and Hitlin and the Pew Research Center (2013) found that the "reaction on Twitter to major political events and policy decisions often differs a great deal from public opinion as measured by surveys." In the analysis of major news events such as presidential debates and the outcome of the 2012 presidential election, Pew found that Twitter posts were often more liberal or conservative than what was indicated in surveys. The study also found that negativity was often overrepresented on Twitter, yet it also could be underrepresented. For instance, "an overwhelming majority (77%) of post-election Twitter comments about the outcome were positive about Obama's victory while just 23% were negative. But a survey of voters in the days following the election found more mixed reactions to the election outcome: 52% said they were happy about Obama's reelection while 45% were unhappy."

Given the kind of incongruences seen in the Pew study when compared to the typical margins of error seen in opinion polls, it is clear that social media analysis alone is not always a reliable means for estimating the current state of a population. This is especially true in populations where technology access is limited or other significant socioeconomic disparities exist. Even in traditional surveys conducted with scientific sampling, non-response can cause substantial bias in opinion estimates; the problem is magnified enormously in social media where only a limited subset of citizens post regularly, and essentially "self-select" into the sample. As such, significant caution should be given to its use in situations requiring critical decision making.

Methodology: A combination of text analysis and simulations

To help extend the accuracy of opinion polls and mitigate the misrepresentations seen in social and news media, this paper discusses a methodology for combining text analysis with simulations of an agent-based model. Our approach begins by initializing the agent-based model directly from polling data to create a synthetic population that is representative of all the respondent's identities (e.g. gender, ethnicity, education, etc.) and initial opinions. This model is then simulated to allow the agents to interact with each other and to respond to the significant events that have been extracted from news and social media. The intended result is then an updated representation over time of the original polling data. In the following sections we describe the simulation model developed, the text analysis techniques employed, and the preliminary results of an ongoing study based on the population of Afghanistan.

Simulating population opinion dynamics with an agent-based model

The original motivation for developing a model of population scale opinion changes was to help decision makers in defense and security environments better understand how local populations affect the outcome of current and future missions. Given the importance of geographic and social distributions, an approach based solely on system dynamics modeling would have required immensely complex subscripts to keep track of the numerous combinations of population cohorts and physical locations. As such, after developing the individual opinion change behaviors in a system dynamics model, the resulting equations were adapted to provide discretized versions within an agent-based model. These versions are then evaluated when an agent receives an opinion from another agent or as one generated in response to a given event.

To create our agent population, we first produced an array of survey respondents that is statically weighted by census data for the regions they live in. After analyzing the array, we determine the minimum number of agents required to ensure that each survey respondent is represented by at least one agent. We then initialize each agent using the identities of a respondent (e.g. gender, ethnicity, education) and by sampling a continuous range that corresponds to his or her response on a Likert scale for each opinion of interest (e.g. local government, key leader, coalition forces, etc.). The overall opinion scale ranges from negative one (strongly opposed) to positive one (strongly for). Each agent is also assigned a distribution of reactions to different event types (e.g. attacks, reconstruction, elections, etc.) that is based on the unique cohort they belong to (e.g. male, Pashtun, uneducated).

After initializing the agents, they are then distributed over geographic regions and connected to one another using the distributions within crosstabs that estimate how cohorts interact and their level of access to long distance communications and media. In more general terms, agents that share similar identity characteristics are more likely to interact with one another and update their opinions to match one another than agents with different identity characteristics. With the synthetic population initialized we are now ready to simulate the population changes over a period of time, allowing them to interact at regular intervals and to respond to events that occur throughout the period.

As mentioned earlier, the process by which each agent updates its opinion in response to the signal sent by another agent or a given event was first developed in a system dynamics model, which is shown in Figure 1 and whose equations can be found in the supplemental material submitted along with this paper. This model was based on an extensive literature review of different social identity and social influence theories as described in Grier et al. (2008). The product of this study was an algorithmic interpretation of the theory of bounded confidence, where changes in an individual's **opinion** are moderated by his or her level of **certainty**. Depending on the disparity between two individuals and the level of certainty held by the receiver of a given opinion, a variety of outcomes are possible.

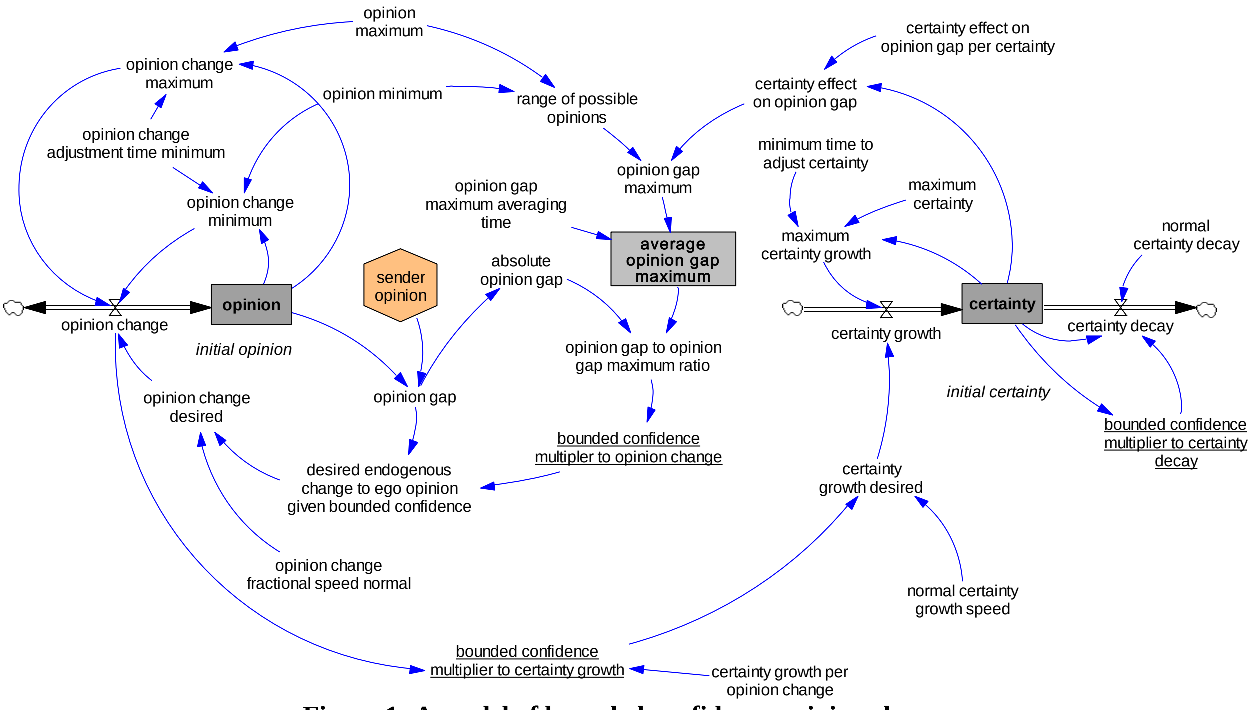
opinion maximum
opinion change maximum
opinion minimum
range of possible opinions
certainty effect on opinion gap per certainty
certainty effect on opinion gap
opinion change adjustment time minimum
opinion change minimum
opinion gap maximum
minimum time to adjust certainty
maximum certainty
opinion gap maximum averaging time
average opinion gap maximum
maximum certainty growth
normal certainty decay
sender opinion
absolute opinion gap
opinion
opinion change
initial opinion
certainty
certainty growth
certainty decay
opinion gap to opinion gap maximum ratio
opinion change desired
opinion gap
initial certainty
bounded confidence multiplier to certainty decay
bounded confidence multipler to opinion change
desired endogenous change to ego opinion given bounded confidence
certainty growth desired
opinion change fractional speed normal
normal certainty growth speed
bounded confidence multiplier to certainty growth
certainty growth per opinion change

Figure 1: A model of bounded confidence opinion change

As shown in the plot in Figure 2, given a sender agent's opinion held constant at 0.25, the amount the receiver agent is willing to close the gap is dependent upon its certainty and the size of the gap between them. With no starting certainty and an absolute gap of 0.5 resulting from the receiver's initial opinion of -0.25, over time, the agent is able to close the gap completely. With a level of moderate certainty (0.5), the agent will close only part of the gap before the growth of its certainty prevents further changes. When certainty is high or the gap is too large, then the agent will not close any portion of the gap, no matter how long they receive the same signal from the sending agent.

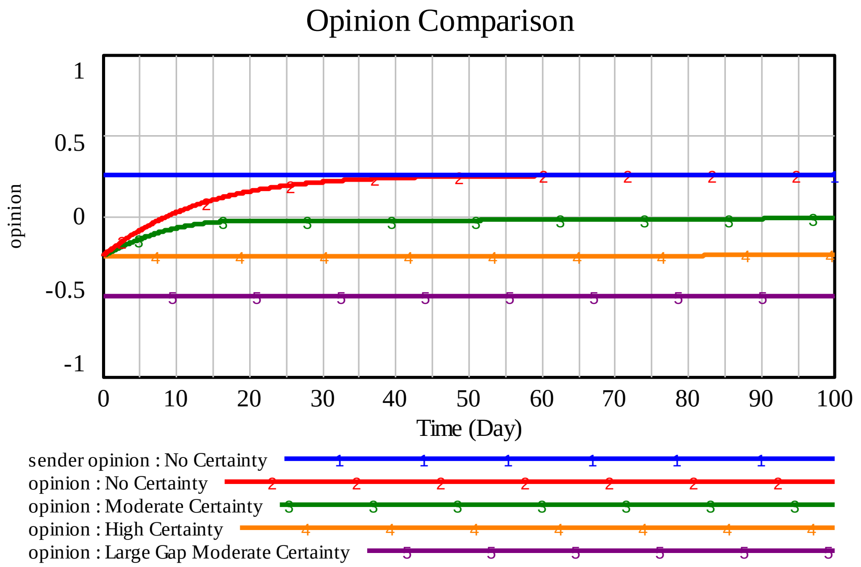

Figure 2: Simulations of opinion change under a variety of bounded confidence conditions

In the agent simulation, the signal received from a sender agent within its network or in response to a given event is not held constant as it is in the system dynamics model, but rather dependent upon discrete interactions. These interactions occur at different intervals that are scheduled based on a distribution of contact frequency. Given repeated interactions between two particular agents and no others, the same general behaviors can be created, but since each agent has a diverse network and because there are external events influencing the agent population, the overall dynamics have a variety of the emergent features that are characteristic of agent-based models. For this reason, we run Monte Carlo simulations on the various parameters affecting initialization and opinion changes in order to estimate the sensitivity and statistical significance of any population scale trends produced by the model.

A model for analyzing the unstructured text

To extract significant events from unstructured text, we use a variety of preprocessing and statistical techniques. One of the important preprocessing stages is the recognition of location names. To complete this, we utilized the GeoNames (geonames.org) database of place names to extract locations referred to in documents. These locations were then used as metadata in the processing of text to create a Dirichlet-multinomial regression (DMR) model (Mimno and McCallum, 2012) of a corpus compiled from the Afghanistan News Center (www.afghanistannewscenter.com) archive of articles. DMR is an unsupervised technique for extracting a set of topics that are conditional on metadata, such as a region in this case. Topics, in this sense, are distributions of words that represent a central concept. For example, one topic of interest extracted from this dataset represents the concept of “civilian casualties” In

Table 1, we present the top words associated with this topic.

Table 1: Features (i.e., words or phrases) statistically associated with a "Civilian Casualties" topic

Topic 36 “Civilian Casualties” Feature	 Probability
civilian	0.0835
casualty	0.0297
air	0.0189
death	0.0177
strike	0.0167
killed	0.0150
incident	0.0141
report	0.0104
operation	0.0103
“civilian casualty”	0.0095

For a given topic, we then set thresholds for the temporal probability (also referred to as the topic prevalence) in order to automatically determine when an event is likely occurring. Figure 3 provides an example of several topic streams with high and low thresholds that are based on the first and second standard deviations from the mean. When a topic moves above a threshold we note the start of an event and then set the duration once the topic prevalence drops below the threshold again. Events that go above the higher threshold denote more significant events that have stronger reaction distributions associated with them. The resulting stream of events then becomes the stimulus for the agent-based model described in the previous section.

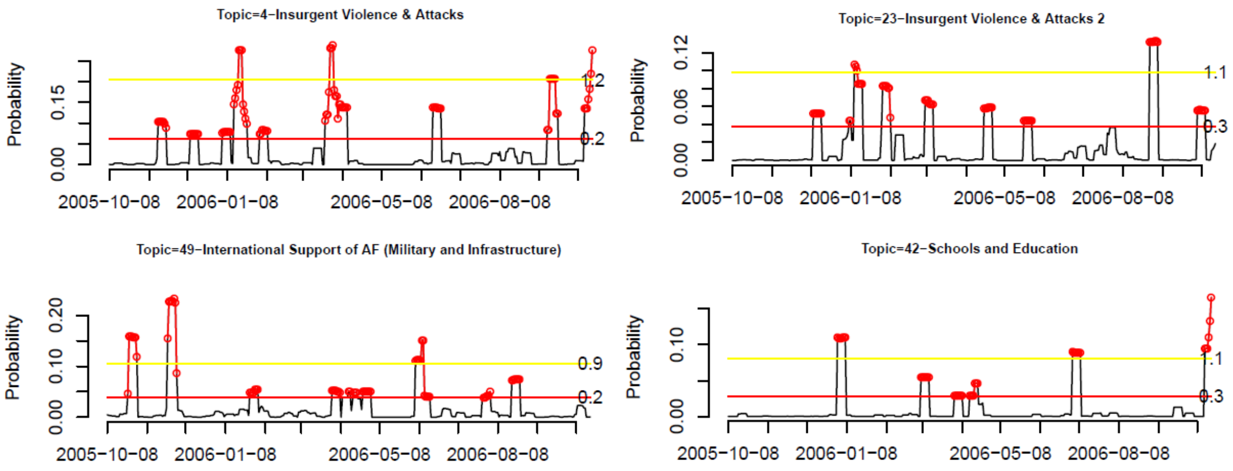

Figure 3: Generation of events based on topic prevalence over time

Preliminary results

To evaluate our approach we are using a collection of opinion polls conducted in Afghanistan starting in 2005 and running through 2008 that were provided by our partners on this research, Charney Research. As discussed in the previous section we have also collected and analyzed news articles covering the same time span from the Afghan News Center archives. At the time of this writing we have just begun assessing the performance of the model over full year intervals. The first interval that is being analyzed spans from 2005 to 2006. The charts in Figure 4 provide one example of the opinions of Afghans towards the Taliban, starting with the model initialization in 2005 and ending with the simulated changes in 2006. In this preliminary assessment, there are no significant discrepancies between the actual and simulated distributions, yet the insights gained from running the study have provided significant opportunities for improvement.

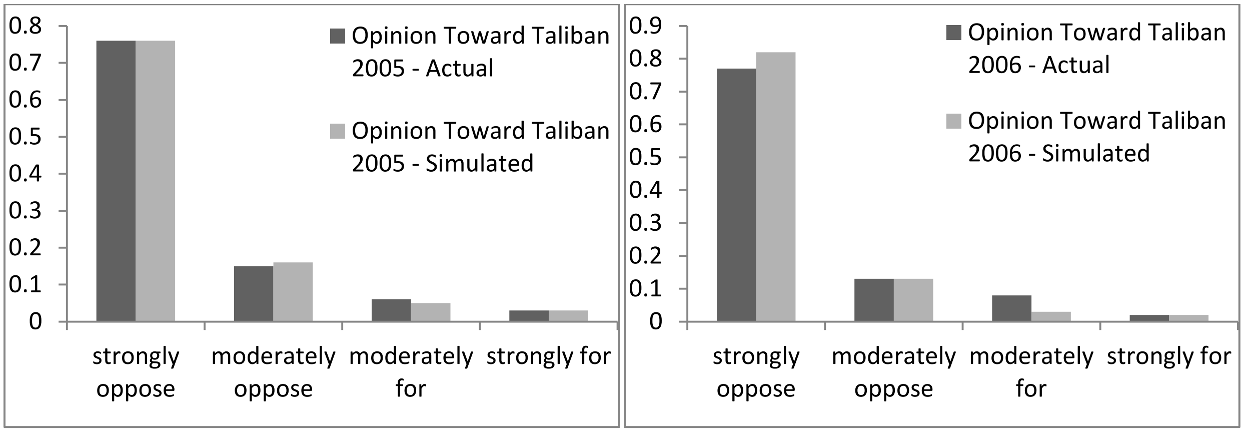

Figure 4: Actual and simulated opinion toward the Taliban

Figure 5 provides a more clear depiction of some of the discrepancies observed, with the shift in opinion towards "strongly oppose" noticeably overrepresented and with the opposite seen in the case of "moderately oppose." Further analysis of the cohorts is also required to see if certain groups performed better or worse than others. These insights will then help to refine reaction distributions and illuminate potential gaps in the events extracted from media and used to stimulate the model.

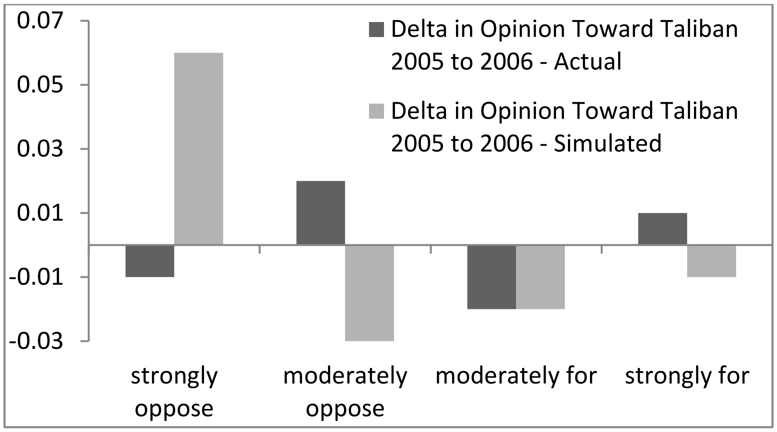

Figure 5: Actual and simulated opinion change toward the Taliban

Conclusions

The intention of this paper was to provide an overview of a new methodology for combining an analytical model of social and news media with an agent-based simulation model of opinion dynamics based on polling data. Given the preliminary state of our evaluation, it is unreasonable to draw any firm conclusions from this work. However, it is clear that traditional surveys and analysis of social media will be insufficient in isolation in accurately representing the state of local population given today's fast changing world. As such, methods like those discussed in this paper will be needed in order to help fill a critical gap in the opinion research community.

References

Asur, Sitaram, and Bernardo A. Huberman (2010). "Predicting the future with social media." arXiv preprint arXiv:1003.5699

Corley, C., Armin R. Mikler, Karan P. Singh, and Diane J. Cook (2009). "Monitoring influenza trends through mining social media." In *International Conference on Bioinformatics & Computational Biology*, pp. 340-346..

Golbeck, Jennifer, Cristina Robles, and Karen Turner (2011). "Predicting personality with social media." In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, pp. 253-262. ACM, 2011.

Grier, R.A., Skarin, B., Lubyansky, A., & Wolpert, L. (2008). SCIPR: A Computational Model to Simulate Cultural Identities for Predicting Reactions to Events. Second International Conference on Computational Cultural Dynamics, September 2008. College Park, MD.

Mimno, David, and Andrew McCallum (2012). "Topic models conditioned on arbitrary features with dirichlet-multinomial regression." *arXiv preprint arXiv:1206.3278*.

Mitchell, Amy and Paul Hitlin (2013). "Twitter Reaction to Events Often at Odds with Overall Public Opinion." Pew Research Center, March 4.