Validating Social Data For Public Opinion Polling

March 2013

www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116 USA

Introduction

Traditionally, polling is a time-consuming and expensive process. We will demonstrate social data processed by machine using sophisticated software can give comparable results in less time, much more cost effectively and with data from millions of individuals, rather than hundreds to thousands.

For comparison, we will be looking at data from a recently-published (March 4, 2013) article from Pew Research Center, titled *Twitter Reaction to Events Often at Odds with Overall Public Opinion*[1], which compared Twitter data against several of Pew's national polls. This article argues that Twitter does not strongly correlate with national opinion polls, and often expresses a strong inherent bias.

We will contend that while Twitter content may demonstrate a bias in comparison with national polling, the social web as a whole can provide strongly correlated results. For this analysis, we will use Infegy Atlas, which is backed by a broad social dataset covering more than 6 years of dialog from numerous online channels. Using robust data normalization, powerful filtering and sophisticated analysis algorithms, we will look for a correlation between traditional polling methods and the broader social web.

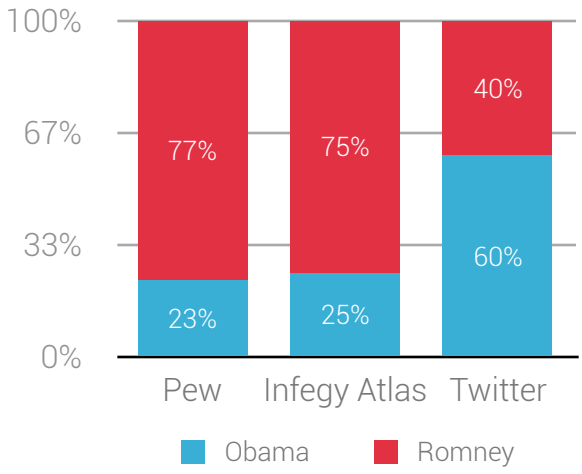
The social web can provide strongly correlated results

Our analysis will focus on eight politically focused polls performed by Pew over the last year. The polls are divided into two distinct groups. First are polls covering general opinion of a subject or event with an undecided outcome, such as measuring favorability of a speech. The polls in the second group measure reaction to events with decided outcomes, such as reaction to an election or supreme court decision.

We hypothesize broad social data will provide a strong correlation to the first group, general opinion polls, and a less strong correlation with the second, reaction polls. We base this on historical findings indicating that in discussions of events which have a decided winner, supporters of the winning side tend to be more vocal of the victory than the supporters of the defeated side.

[1] http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/

Opinions of Undecided Subjects



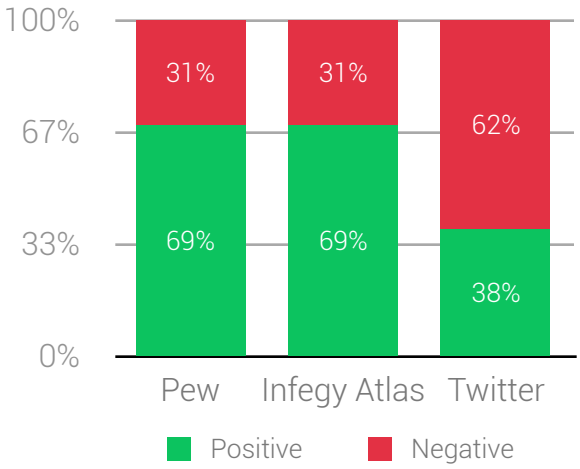
Better Job in First Presidential Debate (Oct 2012)

To start, we examine opinion around the results of the first presidential debate of 2012. The original opinion poll asks which of the two 2012 U.S. presidential candidates were perceived to have performed better. This is an example of an event with an undecided outcome: The result is purely an opinion of individual polled.

For Infegy Atlas' analysis, we've looked at volume of commentary favoring Mitt Romney versus Barack Obama. In this case, our results nearly mirror Pew's poll, with the only variance falling within the margin of error. The Twitter result, however, shows a very different picture, with a much stronger bias towards Obama than the general public.

Obama's Second Inaugural Speech (Jan 2013)

The next opinion poll examines reception to Obama's second inaugural speech. Again, this is not a decided outcome, but rather a measurement of the general opinions of a group of individuals, in this case asking whether or not they liked the speech.



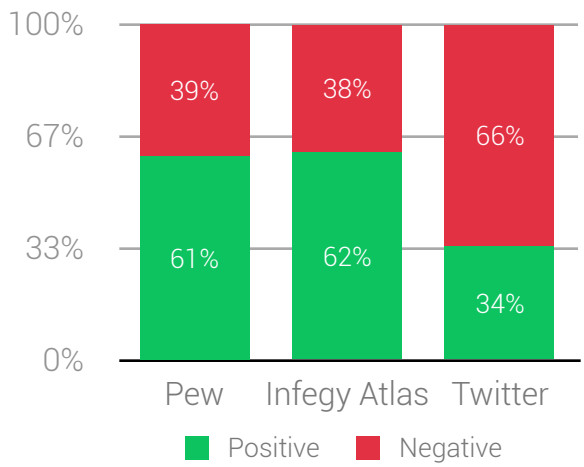
In this case, Infegy Atlas' result comes from an automated sentiment analysis of the social commentary about the speech, measuring favorability of opinion. Again, results closely correlate with Pew's poll, in this case returning identical figures of 69% positivity for those with an expressed opinion.The results from the Twitter-centric analysis are again mismatched, skewing largely negative.

infegy

www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116

Opinions of Undecided Subjects

2012 State of the Union Address (Jan 2012)

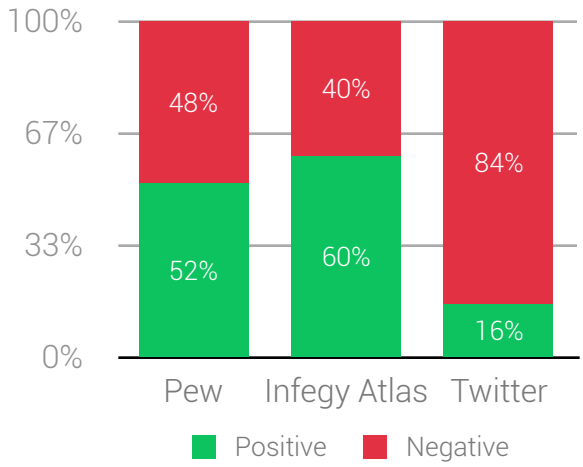


Our third result examines Obama's January 2012 State of the Union address, measuring public favorability of the speech.

This result uses Infegy Atlas' sentiment analysis, measuring expressed favorability. Again our result nearly mirrors Pew's, falling within the poll's margin of error, and again significantly outperforming the Twitter result.

Opinion of John Kerry Post-Nomination (Dec 2012)

On December 21, 2012, John Kerry was nominated for U.S. Secretary of State. This Pew poll is not measuring response to the nomination, but rather favorability towards Mr. Kerry just after his nomination. For our result, Infegy Atlas analyzed sentiment for Mr. Kerry just after his nomination, similarly to the poll.



Here our analysis less perfectly matches the Pew data, though is still fairly close. We see a slight positive skew, that manifests only right after the nomination. As we expected in the original hypothesis, the larger discussion volume from Kerry supporters celebrating the nomination offset the more usual volume of detractors. This reaction effect causes a bit more skew from public polls than we'd see examining support of the Secretary at other times, an effect we examine closely in the next section.

The Twitter data here exaggerates this effect and adds a drastic audience bias, showing extreme negativity towards Mr. Kerry.

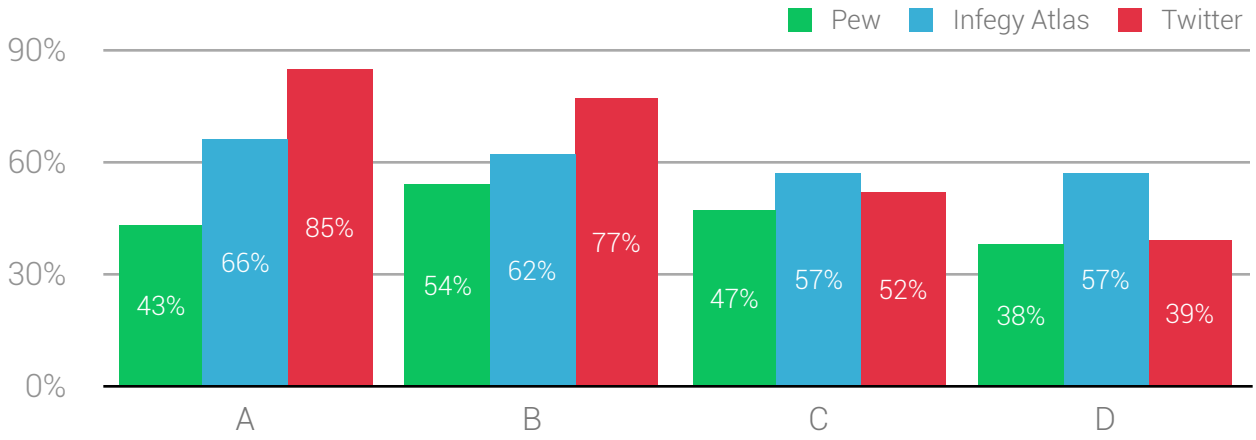
infegy

www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116

Reactions to Decided Events

The second set of polls focus on events with a determined outcome, or officially decided winners. While examining the broad social web, we are measuring public vocal reaction to an event, rather than private general opinion. This is in contrast to a poll, which is able to ask individuals their opinion regardless of whether or not they care to otherwise share it publicly.

Results - Positivity Percentages



A: California same-sex marriage ruling (Feb 2012)
B: Obama's reelection (Nov 2012)
C: Supreme Court health care ruling (June 2012)
D: Mitt Romney's choice of Paul Ryan (Aug 2012)

Result Analysis

As expected, we see a positivity skew in Infegy Atlas' results with these four polls. When the event has been decided, we are measuring publicly expressed reactions to the event. In contrast, the poll results ask the more general population for their private opinions, even those who wouldn't otherwise share it publicly. When an event has this type of conclusive outcome, our data shows the supporters of the winning side tend to be more vocal in response. Those who supported the losing side are less vocal, causing a positivity skew towards the winner.

infegy

www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116

Conclusion

Pew Research Center analyzed results from Twitter against their own polls, concluding there is often a strong perceived bias towards liberalism or conservatism in the Twitter results. When running our analysis, however, we found that a more broad look at online conversation using our own social data demonstrated a much stronger correlation with generalized opinion polling, especially for opinions on something with an undetermined or undecided outcome, such as the sentiment around a speech. While the correlation is less strong for polls measuring reactions to events with decided outcomes, such as election results, the effect demonstrates a consistent bias towards the winner, instead of political or topical bias as seen from Twitter alone.

This is an effect we've noticed repeatedly in our nearly 10 years analyzing this data: People tend to talk more about what they like, rather than what they don't. Given this, when an event has a conclusive outcome, the supporters are more likely to comment in celebration than the opposition will comment in defeat.

Results show strong positive correlations with opinions of undecided outcomes

We have shown here that Infegy Atlas' analysis does an incredible job of gauging public opinion for undecided events. For these events, the average deviation for Infegy Atlas' results against the corresponding Pew polls measures just 2.7%, near the polls' margins of error. We've seen that by analyzing a broad social dataset, in these situations we will capture a relatively balanced view of opinion, giving us the ability to virtually mirror

Deviation against the Pew polls of just 2.7%

traditional polling using social data with a more automated methodology.
This result is quite powerful, as social data can be processed much more rapidly, and returns results more easily than the traditional polling process. This incredible speed and efficiency provides numerous benefits, such as the ability to measure public opinion in near real-time, gauging response even while an event is taking place, and the ability to quickly examine a much larger number of topics in a short time.

Conclusion

So why does the Twitter result vary so much in these cases? While we don't have access to the original Twitter analysis used in the Pew report, our own investigation indicates that analysis focused on Twitter alone will demonstrate strongly skewed results. We have seen that while an ever larger and more representative sample of the population share content on the social web, the individual sources such as Twitter form sub-cultures that skew towards different opinions and interests. Pew's own data confirms this result, showing that broad online demographics[1] more closely match the overall population in contrast with the more skewed demographics seen in Twitter's community[2].

Additionally, with our own analysis, we find Twitter users to be markedly more negative versus the general population. Our analysis of broad political topics shows the overall positivity within Twitter data averages just 39%, versus other channels' more balanced 50%. This further explains the negative bias discovered by the Twitter analysis, and indicates focusing too largely on Twitter, or any other single channel, risks large skewing of a result intended to measure broad public opinion.

Twitter data alone averaging just 39% positivity, versus other channels' more balanced 50%

When gauging general public opinion, population selection is key, whether using traditional polls or online analysis. If the sample population polled does not closely match the general population as a whole, the results can show a distinct bias. Traditional polling methods expend considerable effort to minimize this effect, and it is no less important when examining social data. Much as the results of a poll gauging national opinion based on responses from one town would be called into question, so should measuring general online opinion limited to only one community, such as Twitter. For a more representational sample, the broader social web must be examined. With the broader view, we have concluded that social data paired with a sophisticated software analysis tool such as Infegy Atlas can obtain results highly correlated with the general population and traditional methodology.

[1] http://pewinternet.org/Trend-Data-(Adults)/Whos-Online.aspx
[2] http://pewinternet.org/Reports/2012/Twitter-Use-2012/Findings/Twitter-use.aspx

Methodology

Analysis

All data presented in this report labeled Infegy Atlas has been generated through Infegy's Infegy Atlas platform. Infegy Atlas performs deep analysis of natural language curated from a broad array of channels throughout the internet. Content sources include news media, blogs, social networks, discussion forums and more, and is gathered both in real time and through Infegy's 7+ year historical archive. The dataset is as broad as possible, ensuring analysis includes dialog and opinions by anyone, wherever and however they choose to make themselves heard. All data collected is subject to a sophisticated four-stage filtering process, ensuring only quality, relevant content is used in our analysis and reporting.

Turning our vast collection of dialog into powerful insight takes advanced technology. Developed in-house, Infegy Atlas' cutting-edge linguistics software allows us to truly understand language to extract a wealth of deep knowledge. Powerful technologies understand grammar rules and pronouns, and are subject-specific and self-learning, ensuring the analysis has exceptional accuracy.

For more information on Pew's methodology or that used for Twitter data, please see the Pew report *Twitter Reaction to Events Often at Odds with Overall Public Opinion*[1]. Information on Infegy Atlas queries used and other detail available upon request.

About Infegy

Since 2007, Infegy's cloud-based technologies have been transforming huge volumes of dialog into valuable consumer insights for major brands and agencies. The company's flagship product, Infegy Atlas, is a social media analytics platform that enables teams to truly understand consumers in the social landscape in order to answer why instead of just how many. The platform provides accurate and in-depth measures such as contextual sentiment, passion, topic extraction, thematic categorization, and headline generation. For more information visit: http://infegy.com

[1] http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/

infegy
www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116