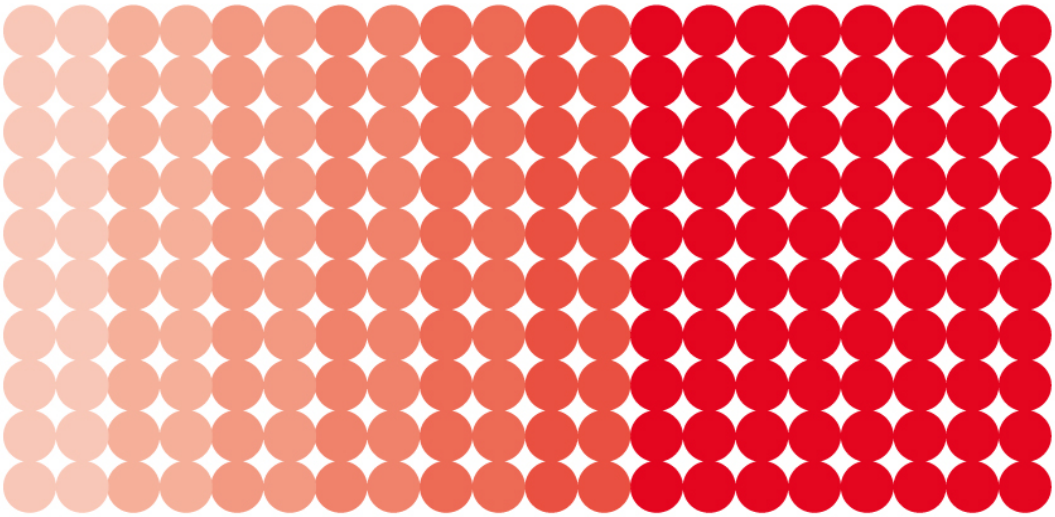


media *policy brief 5*

Semantic Polling
The Ethics of Online Public Opinion

Nick Anstead
London School of Economics and Political Science
Department of Media and Communications

Ben O'Loughlin
Royal Holloway, University of London
Department of Politics and International Relations

OPEN SOCIETY FOUNDATIONS

LSE MEDIA@LSE Department of Media and Communications

POLIS Journalism and Society

LSE Knowledge Transfer: Higher Education Innovation Fund

Acknowledgements

The author would like to thank the two research assistants involved in this project, Yair Levy and Babken Der-Grigorian, as well as the London School of Economics for supporting this research, and Claire Manibog for the visualisation of pollsters results. In addition, we would like to extend our thanks to the various representatives of research firms, political parties, media organisations and regulators who took the time to talk to us.

The LSE Media Policy Project is funded by the Open Society Institute.

LSE Media Policy Project Series Editors
Sally Broughton Micova, Damian Tambini, and Zoetanya Sujon

May 2012.
LSE Media Policy Project. http://blogs.lse.ac.uk/mediapolicyproject/

Key Messages

- As the 2010 UK General Election saw an increased use of social media by politicians, activists, journalists and citizens, a number of consultancy firms began “semantic polling” – the employment of natural language processing technology to “read” and codify vast datasets gathered online, and then the use of this data to illustrate and understand public opinion.

- The semantic polling techniques employed are largely experimental and vary widely between firms. There is also very limited methodological transparency. This is problematic as academic research suggests that statements about public opinion made in media can actually drive as well as reflect popular attitudes.

- Both those carrying out semantic analysis and those in the media reporting it have a responsibility to offer appropriate explanations about the meaning and limitations of the conclusions, and the methods used in data analysis. They can do this by: increasing media literacy among citizens; increasing data literacy among journalists and editors; developing structures for self-regulation; and developing institutions that ensure a greater level of methodological transparency.

Introduction

Following the 2010 UK general election, some commentators downplayed the role of the internet. Their argument was that new communication technology had failed to have a dramatic impact on the election (there was no "Obama effect"). Instead, television, and in particular the Leaders' Debates, was seen to shape the campaign.[1] However, the argument that 2010 was a *television election* downplays a number of important changes driven by new media technology. One of the most important of these is the changing way in which citizens consume and interact with multiple strands of media simultaneously. With citizens watching on a TV screen and commenting on the internet - the two screen phenomenon - new streams of data have been generated.[2] In turn, this creates information that can be used (or at least has the potential to be used) to gain a greater understanding of public opinion. The 2010 election offered an early test for these new techniques. Especially innovative was what we have termed semantic polling techniques. These involved pulling vast datasets from Twitter then machine-reading this content using natural language processing techniques in an effort to quantify the tone of public opinion.

WHY TWITTER?

Of the various forms of social media that have risen to prominence in recent years, it was Twitter that proved to be the most useful resource for those seeking to measure public opinion during the 2010 election.

- *Core design features of the site are brevity and ease of publishing*
- *Individuals post their thoughts and reactions in near-to-real-time*
- *It is a relatively open platform, with most content being in the public domain*
- *Its architecture is built to make this data readily available on a commercial basis.*

For anyone with an electoral, journalistic, commercial or academic interest in public opinion, this is a very exciting development. However, it also requires us to address some questions as a matter of urgency.

Changes in communication technology have always driven changes in the way that elections are fought. In tandem, these developments require a constant reappraisal of how electoral politics is regulated, so that both the legal and informal framework that governs political and civic processes keeps pace with political practice.

.

1. Uses of Social Media Data[3]

Broadly, social media data was used in three distinct ways to make statements during the 2010 campaign.

Vox Pop: Individual tweets were used to provide anecdotal evidence to illustrate particular points being made by journalists. This is a form of electronic vox pop, similar to (and, by implication, subject to the same weaknesses) as street interviews with members of the general public. This technique was particularly in evidence on BBC Newsnight, which dedicated a significant proportion of three post-Leaders' Debate programmes to looking at social media reactions to the broadcasts, employing the website Twitterfall to gather data.[4]

Quantitative Illustration: In this manner of use coverage made quantitative statements about public use of social media. For example, Newsnight employed this second technique for linking social media to public opinion, which involved simple quantitative data. This included citing the number and frequency of tweets, and the number of people producing this content. Additionally, it meant referencing trending topics on the site through hashtags such as #IagreewithNick or #NickCleggsfault with the implication that it reflected a level of public support for Clegg.

Semantic polling: This is mining and natural language reading of textual data such as Tweets to draw conclusions about public opinion and reaction. Natural language processing uses computer technology to read and attribute meaning to textual information, such as whether the author felt positive, negative or neutral about the subject. Huge amounts of data can be coded very rapidly, and broad statements about macro-level reaction can be made.

"We've been asking a firm called Lexalytics to give us instant readouts of Twitter sentiment - last night's chart showed David Cameron [+1.0] and Nick Clegg [+1.1] both attracting roughly equal numbers of positive comments, with Gordon Brown [-1.0] in negative territory."

(Rory Cellan-Jones, BBC, 2010)

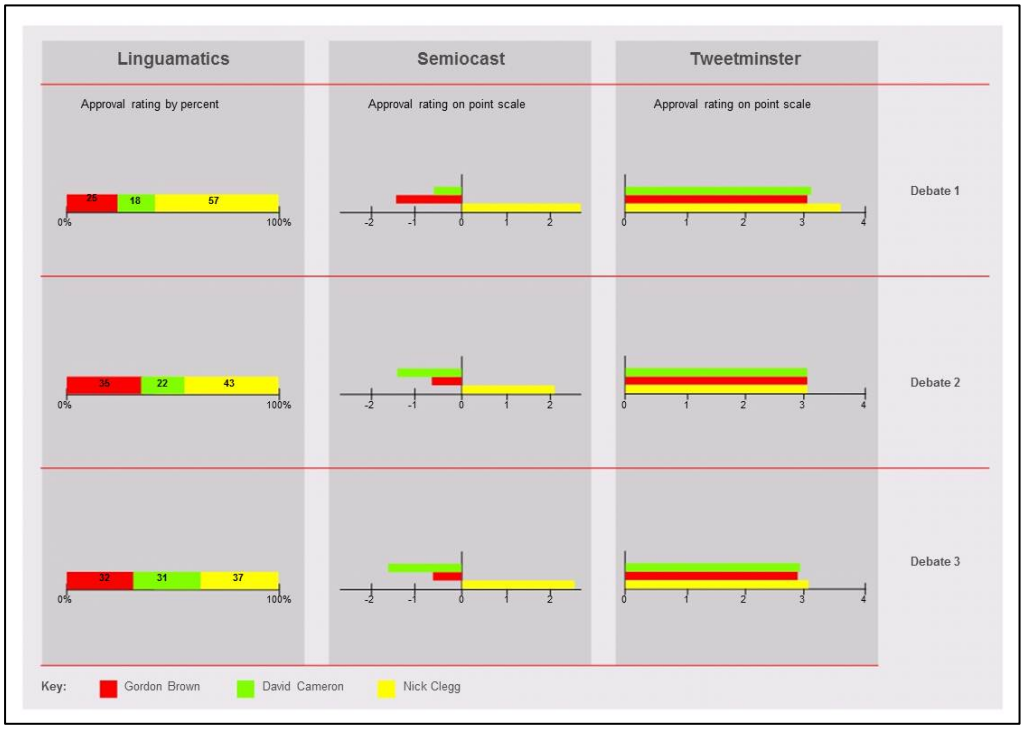
For example, in a blog entry posted after the third leaders debate, BBC Technology correspondent Rory Cellan-Jones cited semantic analysis carried out by the firm Lexalytics, designed to rate the positive and negative sentiment directed at each speaker. As such, semantic polling is slowly creeping into the political and media mainstream.

2. The Semantic Pollsters

A number of companies, including the British Linguamatics and the French Semiocast undertook a similar analysis of Twitter data. London-based political specialists Tweetminster also published Twitter-derived data, while Meltwater Buzz employed semantic technology to measure sentiment across a broader range of social media sources in addition to Twitter, including blogs and other social networks. Cellan-Jones' focus on semantic analysis on his BBC blog reflected a small but growing interest in the possibility of the techniques in the media. The Channel 4 news website published regular blog entries by technology correspondent Benjamin Cohen measuring public reaction to the campaign on social media, while the magazine PR Week employed Meltwater Buzz technology to write blog entries discussing the changing mood of public opinion.

What did these various studies of the same debates present to the public?

Figure 1: Data produced by Linguamatics, Semiocast and Tweetminster following the three leadership debates.



Note the different results produced by the three firms, even down to the numeric technique employed to display it, despite the similar source material they were working with - comments published on Twitter, during the debate

and deemed to be about it. This points to distinctive methodologies employed by the various firms gathering and analysing data.

3. Methodological Differences in Semantic Polling

One of the main reasons semantic polling is currently problematic is that firms use different methodologies, leading to distinctive samples and outcomes. Partially, this related to how they construct their samples. Various search terms might be used - leaders and party names being the obvious examples – but these techniques were capable of generating both false positives and missing content, since, for example, words such as "brown" might appear in non-election related tweets or, for that matter, some political comments might not use the designated keywords.[5]

Once the sample has been gathered, natural language processing techniques are not equal in their sophistication. For example:

> *[T]here are two families of tools for sentiment analysis: one of them is based on lists of words, which are supposed to be positive or negative or neutral… The other family of tools is based on machine learning, which is what we used now and what we used at the time in the UK elections. We took a sample of messages about the leaders, before the TV debate. Manually, we used a tool to say "this message is positive/negative/neutral" and then, the computer was able to analyse during the debate, on the messages that were sent during the debate (data analyst, interviewed for this project, 2010).*

Another methodological difference is to be found in whose Tweets were analysed. Twitter is not an environment of equals. Some people have ten followers, others millions. Clearly then, some tweets have at least the potential to have more of an impact than others. Some, but not all, semantic pollsters build this factor into their methodology:

> *[We were] identifying who those influencers were… so individuals, journalists, news sources, politicians etc that people pay most attention to. And by attention we look at the links they were sharing, retweeting their messages, mentioning their comments and things like that (data analyst, interviewed for project, 2010).*

Semantic polling techniques are still at a very embryonic stage, and this makes it even more important that the methods and limitations of the data are properly understood by journalists and editors, and in turn, explained to the public.

This is especially important if we consider the broad range of academic evidence that suggests that statements about public opinion made by the media have the ability to drive, as well as reflect popular sentiment. Although it is not uncontroversial, many scholars have argued that the publication of opinion polls can create a bandwagon effect, where a publicised increase in support for a candidate, party or policy can lead to more citizens gravitating to that position. Others have argued for the reverse - an underdog effect, where support moves the other way. The way polls are presented, even down to the graphics used, can also have an effect. This pattern is not just applicable to opinion polls, as recent research has also shown that the path of "the worm" (a graphical device showing focus group reaction to televised debates in real time) has the ability to influence viewers perceptions of who "won" the contest.[6]

4. Key challenges raised by the emergence of semantic polling.

The usefulness of social media data

Many people in politics and the media remain sceptical about using social media to measure public opinion. Central critiques focus on the demographic problem with using this kind of data quantitatively. For example:

> *Traditional methods of public opinion polling are very, very cautious about making sure that your groups… are balanced across demographics… I think that's the issue potentially with social media, which is, there are certain types of demographics that congregate around different sites (party campaigner, interviewed for this project, 2011).*

Another of the critiques from people in media and politics is the lack of representativeness found in semantic polling. This critique was also evident during the course of the 2010 election campaign, with some journalists arguing that Twitter was receiving disproportionate coverage, given the relatively limited number of people using it and the narrow socio-economic make-up of that group. This view was echoed among those interviewed:

> *Reaching a situation in which somebody says they contacted 100,000 people through social media, but what if those 100,000 are in the same city or have the same demographic make-up - it's not going to be representative (traditional pollster, interviewed for this project, 2011).*

However, the view that the unrepresentativeness of Tweeters makes the data they produce meaningless was strongly opposed (perhaps unsurprisingly) by those who worked in the data analysis industry. What was most interesting about their argument however was how they reframed the debate, suggesting that the data needs to be thought about in different terms:

> *[P]eople who use more traditional tools would say this is not representative of the voter, consumer, market or whatever. But you can get much more interesting insights by looking into what people talk about, the topics they mentioned and how much they remember… You get more, I would say qualitative data (data analyst, interviewed for this project, 2011).*

One of the advantages of semantic polling is the organic nature of social media analysis – grounded as it is in data pulled from citizens' real interactions with each other, as opposed to being elicited by artificial stimuli such as interview or focus group questions.

> *Social media is very different [from polling] in the sense that you don't ask questions, you look at the data and you look at the trends and try to make sense of that data, so it can compliment and in some sense can be more powerful but you can find trends and insights which you wouldn't have found otherwise because you didn't ask that question. And representativeness also becomes less relevant because you're not trying to measure public opinion as polling data. What you're trying to do is try to understand what influences public opinion, which news article, the opinion of which people, so it kind of tells you the reason why you see certain data in a poll, fundamentally (data analyst, interviewed for this project, 2011).*

Semantic polling should be treated more like focus group data (in that it is qualitative) but it also has one distinct advantage: the data is naturally occurring in the real world, not artificially created by designed questions and exercises that occur in a focus group. This has important implications for how the data is interpreted, understood and presented.

The future of semantic polling

Techniques for semantic polling are going to continue to develop. In addition to basic things like the quantities of data being examined and the speed with which it can be processed, semantic pollsters will be looking to work on the subtleties, especially in coping with things like sarcasm and irony.

There is also a lot more data, both qualitative and quantitative that can be used. Other future developments are likely to be based on integrating other datasets with semantic analysis. This will make it possible to weight data in order to make it more accurately reflect the population as a whole or conduct analysis on segments of the population. This development is especially likely because people place vast amounts of information about themselves into online spaces, all of which could be useful for categorising their statements in a broader political and social context.

Semantic polling is not necessarily qualitative, but rather sits between quantitative and qualitative methods.

Some companies are uniquely positioned to take advantage of such developments. For example, the 2012 United States Presidential election saw Facebook moving into the semantic analysis arena by entering into an arrangement with the news website Politico.[7] While the techniques practised in the Facebook / Politico alliance are relatively rudimentary (a simple positive vs. negative machine reading of status updates, similar to the work done in the UK in 2010), the potential of services like Facebook to generate data on public opinion is vast and profound. This is not simply because of the status updates that users publish, but the additional information social networks hold about them - they are a repository for likes, dislikes, political and religious beliefs, and educational and employment history, as well as, of course, information on their friends, colleagues and acquaintances. This information could be used to scale and segment the users and make far more accurate statements about their beliefs and the social, cultural and economic factors that produce, reinforce or alter them.

The real potential of semantic analysis is only to be partially found in the texts that users published, but also in the ability to overlay various datasets, either extracted from social networks or elsewhere.

Ethics and regulatory challenges

Given these existing practices and potential developments, there are two particular types of challenges facing regulators:

1. Micro-level concerns about individuals, especially related to privacy.
2. Macro-level concerns about the impact of semantic analysis on the course of election campaigns.

The UK has an institutional blind spot in its electoral regulation, with little thought to issues related to social media and potential future uses. This is perhaps not surprising since the UK, in comparison with many countries has tended to favour relatively light touch regulation of electoral politics.

The UK has an institutional blind spot in its electoral regulation.

Perhaps the strongest regulatory instruments related to sentiment analysis are actually found in general legislation, notably the *Data Protection Act*. This could be an especially effective bulwark against the integration of multiple datasets. Indeed, relatively strong legal protection of citizens' information is one of the reasons why data segmentation is a less advanced political practice in the UK, as compared to the US. However, even the Data Protection Act becomes problematic when dealing with closed social environments which hold vast amounts of data on individuals, such as Facebook. While people willingly give up these details, it remains highly questionable whether there was genuine informed consent for the various uses their data might be put to once uploaded.

The nature of semantic polling makes regulation complex. The technique is far less geographically grounded than other forms of public opinion research. As such, it can easily be carried out beyond the reach of national-level jurisdiction.

While concerns for both individual privacy and integrity of the electoral process suggest the need for regulation, it is also important that such new rules do not stifle the potential for the development of research techniques that could benefit the public.

A better approach than government regulation may be the model employed by opinion pollsters, who engage in self-regulation through the British Polling Council. It may be that sentiment analysis gradually moves towards this approach, as it becomes more integrated with the traditional polling industry. For example:

> *[W]e set up groups where we follow some of our panellists on Twitter, this way we know about them and what they do. But it's in early stages for us in this form of research. We certainly see it as a*

> *compliment to other things that we do (traditional pollster, interviewed for this project, 2011)*

If traditional pollsters do become involved in semantic polling, they will likely import their industry norms and practices with them. Some pollsters are also talking about forming partnerships with social media monitoring firms, in order to get access to their tools and expertise.

Will the core values of the British Polling Council, notably transparency, become embedded in the semantic analysis industry? This may pose a problem for semantic polling as many of their techniques are proprietary in nature and thus considered industrial secrets. However, there is a price to be paid for being involved in the political process, and transparency of method is surely essential.

Conclusions

The development of social media monitoring has the potential to empower citizens, making politicians more responsive to their wishes and preferences. However, as with opinion polling and focus groups, there is a need to think about how such techniques can be employed ethically and in a manner that is positive for the political process. To that end, we make four recommendations:

- **Increase media literacy among citizens.** Members of the public may be used to consuming news about traditional opinion polls and discussing the trends and likely outcomes with friends and family, particularly during election campaigns. If semantic polling is not necessarily used to predict public opinion, but to understand how public opinion forms and shifts, then citizens will need to adjust their expectations. Complex debates, such as discussions as to whether semantic polling is a quantitative or qualitative technique, and what conclusions those two approaches can actually allow us to draw, are rarely acknowledged in political commentary presented to the public. There needs to be dialogue between citizens, pollsters and journalists about how social media is being monitored, for what purposes, and with what effects.

- **Increase data literacy among journalists and editors**. As semantic polling moves into the journalistic mainstream, regular political journalists, who do not necessarily have an interest in or knowledge of semantics, natural language processing or even Twitter, will be covering this area. This will make many of the issues highlighted in this paper (such as whether social media data should be treated as qualitative or quantitative in nature) even more significant.

- **Developing structures for self-regulation.** In the 2010 election, there was no regulatory structure that specifically related to the measurement of social media data to understand public opinion. In contrast, polling in the United Kingdom is self-regulated by the British Polling Council. While polling firms can still operate without being in the organisation, all the major firms working in the country are members. Semantic analysis firms should look to this model. The British Polling Council can also play a useful role here by facilitating a conversation about what form self-regulation might take. They are in a particularly strong position to do this, due to the deepening relationship between polling and semantic research.

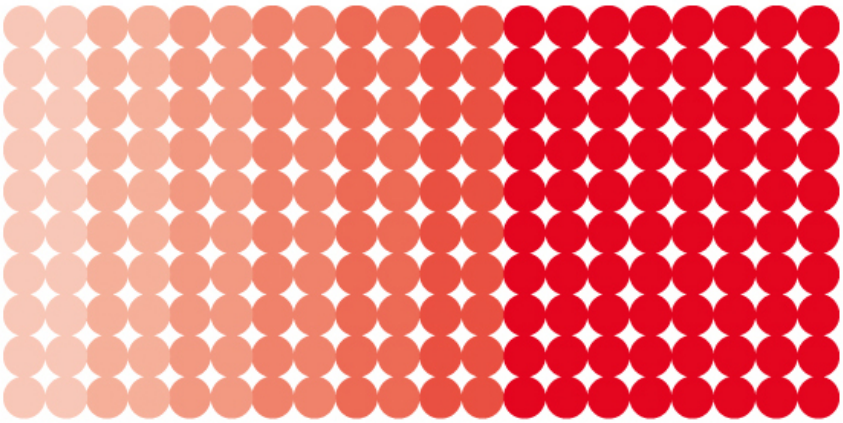
- **Work to ensure a level of transparency**. While transparency is a difficult value to achieve in any technology industry, it becomes increasingly vital if semantic analysis firms are going to continue to publish data on the political process. Entering the political process is a choice which offers companies like those mentioned here benefits, including publicity. In the future though, these benefits must be weighted with a basic requirement of accountability.

"Every political correspondent would need to be a digital political correspondent in the future"

(Journalist, interviewed for this project, 2011).

Notes

[1] Mortimore, R., & Atkinson, S. (2011). Conclusion: Time for a Change? In D. Wring, R. Mortimore, & S. Atkinson (Eds.), Political Communication in Britain: The Leader Debates, the Campaign and the Media in the 2010 General Election (pp. 325-332). Basingstoke: Palgrave Macmillan.
[2] Anstead, N., & O'Loughlin, B. (2011). The Emerging Viewertariat and BBC Question Time: Television Debate and Real-Time Commenting Online. The International Journal of Press/Politics, 16(4), 440-462. doi:10.1177/1940161211415519.
[3] In order to understand this new area of public opinion research, we undertook a set of in-depth interviews. Our sample was drawn from five groups of people with an interested in the emergence of social media monitoring techniques and their application in the political sphere: social media analysts; traditional pollsters;[3] political party campaign officials; election regulators; and journalists who have referred to social media data when writing political stories.[3] In addition to these interviews, we used the data we gathered, plus additional information published during the election to undertake an analysis of the methods employed to link social media data and public opinion measurement.
[4] www.twitterfall.com.
[5] This latter situation was particularly problematic when tweeters were engaged in conversations with each other about a political event. The political word might be used in the first tweet, but not in subsequent elements of the exchange. Some analysts did attempt to cope with this problem by making assumptions as to the likely topic of the reply, based on the original opening of the conversation.
[6] Lang, K., & Lang, G. E. (1984). The Impact of Polls on Public Opinion. The ANNALS of the American Academy of Political and Social Science, 472(1), 129-142. doi:10.1177/0002716284472001012; Marsh, C. (1985). Back on the bandwagon: The effect of opinion polls on public opinion. British Journal of Political Science, 15(1), 51-74; Hollander, B. A. (1994). Newspaper Graphics and Inadvertent Persuasion. Visual Communication Quarterly, 1(1), 8-9; Davis, C. J., Bowers, J. S., & Memon, A. (2011). Social influence in televised election debates: a potential distortion of democracy. PloS one, 6(3)011).
[7] Gannes, L. (2012). Facebook Gives Politico Deep Access to Users' Political Sentiments. AllThingsD. Retrieved January 23, 2012, from http://allthingsd.com/20120112/facebook-gives-politico-deep-access-to-users-political-sentiments/



LSE *media policy project*

About The LSE Media Policy Project aims to establish a deliberative relationship between policy makers, civil society actors, media professionals and relevant media research. We want policy makers to have timely access to the best policy-relevant research and better access to the views of civil society. We also hope to engage the policy community with research on the policy making process itself.

Links Project blog: http://blogs.lse.ac.uk/mediapolicyproject/
Twitter: http://twitter.com/#!/LSEmediapolicy
Facebook: http://on.fb.me/dLN3Ov

Contact Media.policyproject@lse.ac.uk