In the 2010 election, the online space was seen as a battleground to be fought over. In future elections it could be used as a method for better understanding the public.

While journalists speculated about whether the 2010 UK General Election was the country's "first Internet election", semantic polling (using algorithms to read social media data) was under-examined. **Nick Anstead** *and* **Ben O'Loughlin** *explore the role of semantic polling in the 2010 election and argue that it will become even more important in the future.*

We have recently studied how the public reacts to offline events (especially mediated events) using social media. Our first work in this area related to the now infamous appearance of BNP leader Nick Griffin BBC Question Time in October 2009. The second piece focused on social media reactions to opinion polls published in the aftermath of the 2010 UK election Leaders' Debates.

These papers were general in tone, simply trying to document and theorise an emerging phenomeon. However, this got us thinking – would it be possible to extract social media data and make meaningful statements about public opinion from it, in a manner similar to opinion polls or a focus group?

As we soon discovered though, this was not a wholly original idea. Dotted through 2010 election coverage were allusions to the idea that social media did indeed reflect public opinion. Post-debates, Newsnight ran segments on reactions on Twitter, while the BBC's technology correspondent Rory Cellan-Jones wrote a number of blog entries about social media and public opinion. Channel 4 and national newspapers also published this information.

Data from social media in these stories was used in a number of ways. At the simplest level, individual tweets were cited as a sort of e-vox pop. Slightly more systematically, quantative data was used to indicate a high or low level of public engagement with the election, or to show the support for specific politicians through the trending of hashtags such as #IAgreeWithNick or, most famously, #NickCleggsFault.

Most interestingly though, 2010 saw the emergence of a group of firms that engaged in semantic analysis of Twitter. This semantic polling involves using algorithms to "read" tens of thousands of social media items and then coding them according to their content. The data gathered by three firms related to the Leaders' Debate is included in the figure below.

Figure 1: Traditional pollsters and semantic researchers compared, UK General Election debates, 2010

		Traditional pollsters			Semantic researchers			
		YouGov	Populus	ComRes	Linguamatics	Tweetminster	Tweetminster as percentage	Semiocast
Debate 1	Brown	19	17	20	25	3.006	31.09	-1.4
	Cameron	29	22	26	18	3.033	31.33	-0.6
	Clegg	51	61	43	57	3.631	37.55	2.7
Debate 2	Brown	29	23	30	35	3.1	33.33	-0.7
	Cameron	36	37	30	22	3.1	33.33	-1.4
	Clegg	32	36	33	43	3.1	33.33	2.1
Debate 3	Brown	25	27	26	32	2.99	30.51	-0.7
	Cameron	41	38	35	31	2.96	30.20	-1.7
	Clegg	32	38	33	37	3.13	30.20	2.6

For sake of comparison, we have also included polling numbers from three traditional pollsters (we should also add the caveat at this point that this is just a selection of the semantic data published during the election). Of course, this data and the method used to gather it is subject to a number of criticisms. As some commentators noticed at the time, Twitter was an irreverant place in comparison with the starchy seriousness of the debates (and their non-laughing audiences). But can natural language algorithms really cope with irony and sarcasm?

However, perhaps the most obvious issue relates to the type of people who use Twitter. After all, we know they are disproportionately middle class, young, educated and technology literate. Ever since Gallup predicted the results of 1936 US Presidential election, the holy grail fo public opinion research has been representativeness. Is Twitter just a Literary Digest for the modern age?

In the future, that will depend on how semantic research techniques develop. There are three possibilities. The first is that social media data breaks the polling paradigm established by Gallup, and becomes a method more akin to the mass observation, most famously used in the 1940s. As such, representativeness might become less prized and insight into the nuances of how people reason and think could become valued. Second, the passage of time (leading to the normalising of social media use and a population shift) makes social media data more representative. This is, of course, a long term process, although there is some evidence that Twitter is already more representative than it was three or four years ago.

Third is the interesting idea of seeking to apply population segmentation techniques to social media data. The key idea here is interlocking multiple pieces of data. This process is already a big part of the political and commercial world, including pollsters scaling their data to make it representative of the populations a whole and political parties paying a fortune for access to databases such as Mosaic to engage in postcode-based targeting. Think for a second about how much information people put onto social networks – who their friends are, where they work, what they read, and what films, television and music they like (as well as, increasingly, geolocational information). In other words, everything you need to build a complete picture of who they are and where they fit into the national population. If this data could be harvested and overlaid with overtly political information, analysed by natural language processing techniques, it might become possible to create far more sophisticated models of public opinion at given moments.

So we might see 2010 as the embryonic election for this kind of analysis. Indeed, retrospectively, it could seem very innocent, like Harold MacMillan struggling with television (note how he clearly forgets which camera he should be looking at about 1.25 in, and then only realizes after a few seconds). Indeed, if things were to develop along the lines of the third scenario, then a whole host of questions are raised. Do the public really understand what might be happening to information they post online, and the type of picture it could be used to create of them personally? Given that Twitter, Facebook and whatever follows them are corporate actors, what obligations do they have? How open to manipulation is the online space, given that in 2010, many political parties saw it as a battleground to be won, rather than as a method for understanding the public? Who should regulate the way the data is gathered and presented? At the moment, pollsters engage in self-regulation through the British Polling Council. No such body exists for social media analysis.

We are now continuing with the second strand of our research, which involves interviewing a number of political actors from the data campaign of 2010 – party campaign managers, journalists, data consultants, traditional pollsters and election regulators. Our preliminary prediction is this: social media data generated through semantic analysis will be big in the 2012 US election, and integrated in to public opinion studies by the (likely) UK election of 2015.

Read the longer paper, 2010: The Semantic Analysis Election?, here.