Social media metrics: measuring manifestations of the public in online spaces

Can social media help us better understand the public and its reaction to political events? That is a question I first asked myself in 2009, when Nick Griffin, the leader of the far right British National Party (BNP) appeared on BBC Question Time, a flagship news discussion show in the UK. The broadcast caused great controversy, as traditionally politicians of the far right were denied a platform by the state broadcaster. However, the BNP's recent political success – when they won two seats in the June 2009 European Parliamentary Election – was used as a justification by the programme's producers to offer an invite to Griffin.

In the run up to the programme, the merits of this decision were discussed both in the national press and in the political blogosphere. However, it was when the broadcast started that things really began to get interesting. It soon became clear that many people were not just looking at their television screen, but were also absorbed by the potential of a second screen, be it their phone or laptop.

Working with Professor Ben O'Loughlin from Royal Holloway, University of London, we managed (using some home brew code pulled together by my brother, who is a proper computer scientist) to extract about 70,000 updates from Twitter published in the hours around the programme. Looking back, this seems like a relatively innocent time for undertaking such a project, for a couple of reasons. First, and although it seemed like a massive number to us at the time, 70,000 tweets would soon be dwarfed by the vast amounts of data that could be harvested from social media, and second because, at this point, Twitter was still run as a relatively open environment, so the data we required was relatively accessible. It wasn't to be very long though before Twitter attempted to "monetise" the data it had at its disposal, which involved the dual strategy of making the sites application programming interface (API) a lot more secure and then attempting to sell the data to interested parties.

We examined the data we had gathered, leading to the publication of a paper. In it, we made a number of observations as to what viewers were saying as they watched the programme. In particular we noted three things. First, citizens were commenting on what they saw on the screen. In particular, it was notable that they were laughing at Griffin, ridiculing the arguments that he was making. Second, and in addition, viewers were also adding new information, going beyond what was happening on the screen, engaging in what we termed annotation of content. One example of this was the appearance of a photograph of Griffin

attending a National Front rally in the 1970s wearing a white power t-shirt, which went viral during the programme and was re-tweeted hundreds of times. Griffin had achieved much of his political success by adopting a more measured tone than was traditional on the British far-right. His aim in taking this approach was to make racialised politics seem more reasonable and thus more socially acceptable to elements of the electorate that had previously rejected the ideology. However, this old image struck at the heart of Griffin's project, linking him directly to the thuggish and violent street politics that characterised the British far right in the 1970s (it should be noted that, as well as this symbolic value, it also highlighted a fundamental truth that Griffin had been less than honest about – he was heavily involved in the National Front in his youth).

Figure 1: The photograph that went viral on Twitter during BBC Question Time showing Nick Griffin at a National Front demonstration in the 1970s

The idea of annotation suggested that citizens were heavily engaged in the broadcast, to the point where they were undertaking their own relevant research (or at least publishing pre-existing knowledge). However, the third observation we made – salutary for political scientists – was the relatively low levels of engagement among most people. We found many tweets from people who were aware of the Question Time broadcast and maybe even interested in it, but where mentions of the programme were often secondary to the other things that were going on in their lives – going out for the night with friends, helping children with their homework, or cooking dinner.

This dichotomy – between highly engaged citizens adding new information and those for whom politics is at best on their peripheral vision – raises some really fundamental questions about social media data, and how closely we can equate it with the idea of the public. On the one hand, we could suggest that the great weakness of social media is its frivolous, almost non-political nature. As scholars such as [Matthew Hindman](#) have pointed out, only a tiny proportion of what happens online can really be regarded as political. However, drawing on the example of the tangential tweets we found in our Nick Griffin Question Time example, we might also argue that this kind of data has a really special richness, exactly because politics is only a small part of it. As such, it offers a much better chance for us to understand how politics fits into real people's lived existence.

--

We were far from the only people thinking along these lines. Indeed, by the 2010 UK election the idea of the Viewertariat and two-screen viewing was of increasing interest to a variety of groups. There were a number of reasons for this. Traditional media had certainly come closer to embracing social media. When we surveyed the Nick Griffin episode of Question Time, the programme did not even mention Twitter in the broadcast. Rather, viewers were taking it upon themselves to engage with social media. However, by the time of the election, broadcasters seemed far more aware of the potential of social media, making more use of their own accounts and encouraging viewers to comment using pre-designated hash tags. In addition, the penetration of technology had also moved on apace, [with smart phone ownership increasing from 7.2 million handsets in May 2009 to 12.8 million handsets in May 2010 in the UK](#). Similarly, more people were also using these devices to access social networks and publish content. Politically, the landscape was changing too, with the agreement among the three major British political parties to hold the country's first ever Prime Ministerial debates during the course of the election campaign. This combination of factors explains why the 2010 election saw such interest in social media and what people were saying about the campaign.

Developments in other sectors were to prove equally important though. The analysis of the data generated through social media is just one strand of what has been termed the rise of

"big data". Drawing on trends in data gathering and computer-based natural language processing, adherents to big data science argue it has the potential to revolutionise government services, business-consumer relations, disaster management and healthcare provision, to name just a few areas. Related to this, the rise of social media has also had a huge impact on marketing. In theory, firms can monitor the online space to better understand their customers and their own brand, as well as head off any potential scandals that could damage their profitability. A number of specialist firms have recently opened for business, claiming that they can offer their customers insights into how the public perceives them, simply by monitoring social media.

We have been here before – in the 1930s, opinion polling was developed as a tool to better enable marketing, while in the 1990s, focus groups and dial testing emerged to understand consumer reactions to products and the advertisements used to sell them. Both rapidly crossed over into the political sphere, the former most associated with George Gallop's successful prediction of the 1936 Presidential election, the latter with the electoral success of the modernised left under Bill Clinton in the United States and Tony Blair in the United Kingdom. So it should come as no surprise that many of the same firms looking to sell their social media monitoring techniques to the private sector also attempted to the turn their hands to measuring public reactions during the election campaign.

Our research in this area (which is currently under review for journal publication), which involved interviewing a range of people working for social media analytics firms, traditional opinion pollsters and journalists covering social media data, found that many of the firms producing social media data saw the election as an effective way to generate publicity for their businesses. In other words, they did not see what they were doing as being "political" in a meaningful sense of the word, but instead saw it as a corporate activity to boost their profile, with the aim of winning more business. Nonetheless, social media data was crossing over into the media's election coverage. As well as interviews, we also undertook a content analysis of how social media was used to represent the public during election campaign coverage. Our sample drew on a combination of print and broadcast media.

Analysing this data, we were able to find three distinct ways in which social media was used to represent the public.

Figure 2: Anti-Nick Clegg newspaper coverage following the first televised debate. This led directly to the #NickCleggsFault hash tag on Twitter

At the simplest level, individual social media updates were quoted by journalists. These essentially fulfilled the function of electronic vox pops, where the voice of an individual was taken to have some larger representative function, demonstrating a particular strand of public opinion. A second type of social media public was created by the quoting of relatively raw statistical data. This took two forms. Either it involved simply quoting the number of tweets that were published relating to a specific event as an indication of engagement or disengagement (as for example occurred when the number of related-tweets published per minute during the televised debate was widely quoted in post-broadcast coverage) or discussing the rise of election related trending topics on social media. The most notable example of this occurring during the election campaign took place with the rise to prominence of the hash tag #NickCleggsFault. This followed the first televised leaders' debate, which led to a sudden surge in support for the UK's third party the Liberal Democrats and their leader Nick Clegg. In the aftermath of this, Conservative Party supporting newspapers ran a number of stories attacking Clegg and his party. However, these were so

transparently political motivated in nature, Twitter users mocked them by posting updates with the tag #NickCleggsFault. These blamed Clegg for everything from the death of the dinosaurs to rain on someone's wedding day to the entire musical output of David Bowie between 1984 and 1993. Very soon, the mainstream media were reporting this trending topic, creating the type of hybrid media event noted by Andrew Chadwick, wherein old and new media interact to create the news coverage of an event.

The final (and technically most sophisticated) way in which the public were linked to social media is what we have termed semantic polling. The reason for this is that, in part at least (and with many caveats which I will discuss below) replicates elements of the opinion polling method. In particular, semantic polling attempts to quantify a divided public with differing opinions. The starting point for this process is the harvesting vast datasets relating to a specific event. These datasets are then "read" by a computer running natural language processing technology, in order to attribute a sentiment value to them – in other words, in order to quantify whether the social media public has positive or negative feelings towards, a person, party, policy or event.
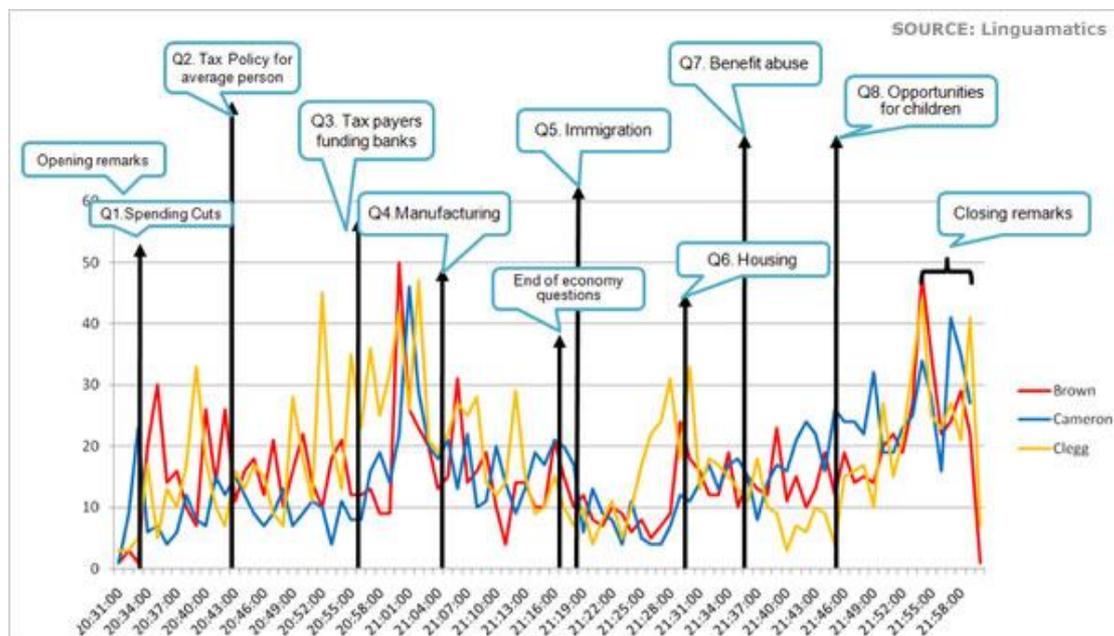

Figure 3: Linguamatics data from the third Leaders' Debate in the 2010 UK election

This type of research was taking place during the 2010 election campaign and also made it into the mainstream media. Notable in the coverage of this type were blog entries written by the BBC's technology correspondent Rory Cellan-Jones (for example here), where he

covered data produced by firms such as [Tweetminster](#), [Linguamatics](#) and [Semiocast](#). This data was often displayed in a graphical format, similar to that shown in Figure 3.

Our interview data showed two ways of feeling towards this development – either excitement or concern. The division (approximately, although not absolutely) tended to exist between those carrying out social media research and traditional public opinion researchers, notably opinion pollsters. The reasons for excitement focused on the potential of these new methods to understand the public's reaction to events faster than was previously possible, potentially even developing techniques for measuring instant response. In addition, advocates of semantic polling argued that their techniques had the potential to offer a richer, deeper understanding of how the public linked up their different political belief with events.

However, in response to this, pollsters offered a very strong critique of the technique. For them, the issue was very clear: a lack of representativeness in the sample. The people commenting on social media were essentially self-selecting and embodied particular structural inequalities, related to access to and use of social media, as well as being disproportionately interested in politics. This, of course, is an anathema to an opinion pollster. Indeed, the very [founding myth](#) of the modern opinion polling industry is constructed around the virtue of representative sampling, as opposed to self-selecting open samples.

So this leads to the question: can these two points of view be reconciled?

--

The development of semantic polling - and its shortcomings - can be understood in two ways, either through an epistemological or an ontological prism. This decision also has important ramifications for how we see the future of the method.

From the epistemological perspective, we might imagine one of two developments. The first of these is to focus on who is publishing social media data. At the moment, as our interviews with opinion pollsters indicated, the types of people who are posting social media updates online about politics are far from typical of the rest of the electorate. They are younger, wealthier and more educated, as well as being disproportionately urban-dwelling. However,

this might not always be the case. Social media use has grown exponentially in recent years, and this trend will likely continue. As more and more people start to use the technology, then the results of semantic polling will likely become more closely aligned with the overall numerics of the population.

An alternative version of this future is to use similar techniques to those found in Internet polling, pioneered by survey firms such as the UK's YouGov. Here, demographic and social categorisations are used to remodel raw data to make it a better fit for overall population. So for example, if the survey results contain too few women over 65, the answers given by members of this group can be more heavily weighted than, for example, those given by men under 30. The type of data available online, where people share information about their education, work, lifestyles and media consumption, certainly ensures that such techniques are possible (indeed, this is one reason why the future of this type of social media monitoring may be dominated by Facebook rather than Twitter. While Twitter presents social media data in a more accessible raw form, Facebook pre-packages many of these key variables as part of a user's profile).

A second epistemological development which may occur would involve a fundamental shift in how the results of social media monitoring data are presented and understood. Instead of seeing big data as a quantitative mechanism for understanding the public like an opinion poll, we could instead think of it as being a giant, always-on focus group providing rich qualitative data. As such, the issues of representativeness becomes a lot less significant. Instead, the data is understood as being useful for understanding the ebb and flow of opinion, and how people understand their opinions in the context of the broader world they live in.

There are reasons to doubt that such a shift is possible, however. As Susan Herbst has documented, quantitative data carries a symbolic authority that qualitative data does not. Furthermore, we know that when qualitative data enters the public sphere is tends to become quantified in some form. For example, research has shown that journalists reporting focus group data have looked to express the information numerically, for example stating how many in the group supported one position or another. This is clearly a misrepresentation of the data, but has an appealing logic in societies where the pubic and public opinion is most commonly expressed through opinion polling data.

There is an alternative ontological approach for better understanding the development of semantic polling. Instead trying to redefine the data, we could instead redefine the concept

that we are looking to understand. In other words, we could rethink our understanding of the public and public opinion, broadening it beyond the paradigmatic definition developed by opinion pollsters.

How might such an exercise work? It is important to note that our modern conception of public opinion is a relatively new invention, dating back to (and gradually achieving dominance since) the 1930s. Prior to this time, very different conceptions of the public had existed. This leads to two important conclusions: first, our conception of public opinion is neither value nor assumption free, and second, it is not immovable.

Opinion polling is modelled on the one-person-one-vote logic that governs electoral politics. As such, it is no coincidence that it came to prominence at the same time as democratic societies were constructing political cultures based on this model of politics. Indeed, George Gallop was unapologetically ideological in justifying the use of opinion polling as a democratic tool, claiming (notably in his 1940 volume *Public Opinion: The Pulse of Democracy*, co-authored with Saul Rae) that the scientific nature of opinion polling returned power to the people and took it away from political elites, who previously might have been able to rhetorically claim the power of public opinion as being on their side.

Yet as some contemporary critics of Gallup's approach, such as American academic Herbert Blumer, pointed out, this approach neglected some of the more complex aspects of public opinion. For example, the methodologically individualist approach taken by pollsters did not factor in the position that a person had within the social hierarchy of a society, and their ability (or lack of ability) to influence the opinions held by other people. Furthermore, Blumer argued, in order to really have an *opinion*, citizens have to be engaged and knowledgeable about a topic, at least to the point where they can debate it with each other. In contrast, opinion polling, where people are simply presented with and asked to choose between a number of pre-decided options generates results that are better described as *mass opinion*, not public opinion (a point perhaps made most forcibly nearly three decades after Blumer's original argument by the famous *Repeal of the Public Affairs Act* paper).

It might be worth going back even further in time for alternative conceptions of public opinion. In the 19th century, for example, political theorist James Bryce used to talk about the "organs of public opinion" in his study The American Commonwealth. These included newspapers and public meetings. Of course, Bryce was writing in an era before modern polling techniques were available. However, the way in which he defines public opinion is

notably broader than most definitions employed today. Furthermore, many of the environments he talks about have similarities with contemporary social media. In particular, they are made up of self-selecting groups of individuals who have chosen to go and offer their opinions.

--

Social media monitoring is clearly here to stay. Big data and natural language processing will continue to be a part of political life, and will likely be used by journalists covering future elections even more. Measured against contemporary techniques for measuring public opinion, notably opinion polls, semantic polling seemingly has some shortcomings. However, a better way of approaching this discussion would be to look at the strengths of the techniques, and how they complement each other. This forces us – academics, public opinion researchers, journalists and the public themselves – to be reflexive about the types of statements we are making and how we understand public opinion data.

And ultimately, this may be the challenge of the whole big data revolution. It will not bring the certainty or easy answers that some of its adherents promise, but instead will bring complexity and nuance. But we should not see this as a bad thing or be fearful of it.