Opinions that matter The hybridization of opinion and reputation measurement in social media listening software

Baptiste Kotras

Opinions that matter

The hybridization of opinion and reputation measurement in social media listening software

Baptiste Kotras (INRAE / LISIS)

Abstract

Since the 2000s, numerous start-ups and agencies have argued for the necessity of analyzing social media data to "know what people think", as they are deemed to provide access to spontaneous expression of thoughts, tastes and representations. How do these actors, and the various types of knowledge and technology they draw upon, change the way we know and act upon people's opinions? This paper offers insight on these understudied actors, by describing the emergence in France of a market for measuring online opinion. It shows two distinct trajectories of innovation, and the key role played by the early clients of these companies and by the demand for tools for online reputation management in the shaping of these instruments, and the definition of epistemic value. Both approaches of online opinion break with the classical egalitarian conception of public opinion. They instead conceive opinion as a mediated and collective process in which not all opinions have an equal value.

Keywords

Introduction

As early as 1998, James R. Beniger, president of the American Association for Public Opinion Research, had the idea that the internet could profoundly renew the way pollsters study '*how we form our attitudes and opinions, are influenced by mass media and by each other, make decisions as consumers and as voters*' (Beniger, 1998: 443). Twenty years later, this revolution has happened, although his fellow pollsters had almost nothing to do with it, as they were more interested in the spectacular cost reductions allowed by the emergence of online panels. Meanwhile, a billion-dollar industry called "social media listening" (SML) has emerged, which allows brands to buy tools and services to monitor public online discourses on blogs, forums and social media platforms (1). However, these international software vendors, such as Brandwatch, Synthesio or NetBase, almost never assess public opinion on national concerns or elections: they work in the discrete areas of corporate communication and market research. Their expertise relies on a set of tools and practices that show little resemblance with traditional polls and surveys. How did these new intermediaries of the 'data analytics industry' (Beer, 2016b) capture social publications for the benefit of private companies? How do SML firms transform the way in which people's *opinions*, in a broad sense of preferences, judgements and representations (Quéré, 1990), are measured and acted upon in the digital era?

Much like other big data advocates, these companies share a strong epistemological claim. Since the 2000s, numerous start-ups and agencies – some of them having become the large software companies mentioned above – have argued for the necessity of harvesting and analyzing social media data to '*know what people think*', as says the tagline of Brandwatch, one of today's main SML vendors. As online expression has become massive, the considerable amounts of texts we produce everyday provide access to spontaneous expression of thoughts, tastes and representations (Cardon, 2010). These start-ups, along with Beniger's early intuition, deem the web to be a powerful way to overcome one long-standing criticism made to traditional survey-based inquiries, both in matters of public opinion and market research: their artificiality, and their tendency to impose their questions and their answers upon respondents (Blumer, 1948; Bourdieu, 1979). Nonetheless, the abundance of uncontrolled and mostly anonymous online textual material faces actors with a set of new methodological challenges, in order to ensure their epistemic value, as online opinions can hardly be considered as representative of the general population. In this paper, I study the emergence of the SML market, and show how its different players articulated epistemic projects and material resources with the demands of their early clients, and how this process shaped hybrid quantification devices combining the measurement of opinion and reputation.

Despite their growing importance in the routine practices of a wide number of firms, we currently lack investigations about SML tools in the social sciences (2). On this subject, academic contributions seem either too close or too far away from its material devices and practices. On the one hand, numerous studies use social data as an empirical fieldwork, and rely on social media listening technologies – be they locally developed or bought to SML firms – to collect, parse and analyze it (Brooker et al., 2016). These contributions to the now well-established field of digital social sciences have focused on a variety of topics, from early attempts to map hyperlinked communities (Rogers and Marres, 2000; Watts and Dodds, 2007), to re-visitations of the agenda-setting theory (Vargo et al., 2018; Whyte, 2016) and the study of public matters, as with the examination of political preferences (Ceron et al., 2014), and election outcome predictions (Gayo-Avello, 2013). This stream of research, from which SML companies themselves have sometimes borrowed, helps us to understand digital worlds. However, technologies for collecting and

analyzing social data are their method, not their object of study, and they therefore provide little historical and critical account of these specific knowledge instruments.

On the other hand, a wide number of contributions, mainly in the field of surveillance studies, have focused on corporate capture of user-generated data, highlighting several risks in terms of social sorting (Lyon, 2003), market surveillance (Pridmore and Zwick, 2011; Trottier, 2012) and algorithmic governmentality (Rouvroy, 2018). Authors like Shoshana Zuboff (2018) have theorized the concept of "surveillance capitalism", arguing that the continuous capture, extraction and valuation of data is now a key feature of today's capitalist societies, reconfiguring market positions and political power. Yet while it raises important and engaging questions, this line of research often remains based on the study of public controversies raised around hyper-mediatized high-profile cases such as Facebook and Google. It consequently tends to lack in-depth investigation, and to overlook less visible but crucial aspects of the corporate routine use of big data, as in the little-advertised case of social listening software.

In this paper, drawing on social studies of science and the sociology of algorithms, I wish to address these gaps in the literature, by offering a detailed account of the material and epistemic cultures involved in the conception of social media monitoring technologies (Knorr-Cetina, 1999; MacKenzie, 2011). Following David Beer's (2016a) research program, I will examine the articulation between these technologies and specific forms of knowing and acting upon opinion, paying particular attention to the properties of the "calculated publics" that they produce (Gillespie, 2014; Yeung, 2018). To this end, I will retrace the innovation paths followed by pioneers of online opinion measurement during the early days of this market (Boczkowski, 2004; Flichy, 2007), to get a grasp on big data technologies "in the making" (Latour, 2005). This way, I wish to account for their genealogy, the actors and epistemic projects involved in it, and on the choices made by their designers, to understand how they shape the production of knowledge about people's opinion on the internet.

Methods

This article draws upon a five-year investigation conducted for my PhD thesis. More specifically, it relies on twelve in-depth interviews, conducted between 2012 and 2015, with French pioneers of the SML market: start-up founders, analysts and programmers from six different companies, which were involved in the very beginning of SML, in the second half of the 2000s. These interviews lasted between 1h and 2h45; they cover the major French players in this market. They were coded manually with a specific attention to individual histories, start-up trajectories and accounts of technological choices, difficulties or successes.

Focusing on French companies, in this case, offers interesting insights on the global history of these actors and their instruments: the first French start-ups interested in online opinion emerged as early as 2005, a date comparable to their British and American competitors; some of them (Synthesio and Linkfluence) are still today among the main international players in the market (3). Above all, as I will demonstrate here, the French market shows in its early years a great diversity of approaches, with startups drawing from several heterogeneous epistemic backgrounds in social and computer sciences, something that ultimately appeared to be specific to the French case when I then interviewed American players of the SML market. This contribution thus calls for future comparisons with other pioneer countries in SML, such as the USA, Canada or the United Kingdom.

The first two sections respectively describe the two main trajectories of innovation in social listening, rooted in heterogeneous material and epistemic resources, and the key role played by

early corporate clients of SML companies in the shaping of their technologies. The third section discusses the convergence of these trajectories to produce reputation management devices, which hybridize public opinion measurement with the tradition of public relations and news monitoring. The conclusion offers perspective on how big data transforms established social quantification and suggests future lines of research.

Sampling influencers, from politics to digital public relations

One first path of innovation involves a group of companies created by people from humanities, marketing and communication backgrounds. These agencies and self-declared '*web institutes*' share a similar history and business model: they were born in the 2005-2007 French electoral context, and began working in the political sphere; their analysts perform both opinion analysis and recommendations. Their little technological resources, and large companies' sudden preoccupation for their online reputation, in the mid-2000s, oriented their epistemology towards the careful selection of digital influencers, for the needs of corporate communication and public relations (PR).

Politics as a springboard for experimentations

This first approach to online opinion is born from various experimentations developed by relative outsiders in political staffs in the mid-2000s, a situation comparable to the future founders of Blue State Digital during the Howard Dean campaign of 2004 (Kreiss, 2012). One key character here is the entrepreneur Benoit Thieulin, known for his role of digital adviser in Ségolène Royal's presidential campaign (2007) and other responsibilities in the public sphere. The French Government Information Service (GIS) hired him in 2000, initially to coordinate the conception of the internet pages of the various ministries. He then set up, in 2003, an experimental observatory of online public opinion, based on the manual review and analysis of press comments and political blogs, whose number was soaring at the time (Siles, 2017). He aimed to identify trends in online debates about the upcoming referendum on the Treaty establishing a Constitution for Europe (TCE). In an environment largely dominated by traditional pollsters, B. Thieulin justified his idea based on the qualitative aspect of his method.

> I had no representativeness, so of course my thing quickly became: ok, I understand that in quantitative terms, I can't say anything […]. On the other hand, I can quite easily observe how an argument is constructed, how opinion crystallizes. And I see that, when we argue that 'Europe is for peace', people answer: Stop kidding me, just because I'm going to vote no, we're not going to declare war on Germany!" – B. Thieulin (interview, August 2012).

Although his method could not predict electoral results, argued Thieulin, it provided nonetheless an otherwise unreachable insight on the production of voting intentions and on the arguments being discussed. Such an insight was based on the possibility of analyzing spontaneous discourse: '*The Internet is an open-air environment, contrary to how pollsters construct opinion. It's open-air, it's real-time, it's life-size*' (interview, August 2012). Along with now widespread statements about big data (Anderson, 2008; Mayer-Schönberger and Cukier, 2013), these easily-accessible, real-time and spontaneous data were considered to be the means of a whole new – and better – way to measure public opinion. However, Benoit Thieulin, trained in economics, lacked the technological support he felt necessary to keep on with this experimental observatory. In 2006, he convinced his superiors at the GIS to buy services from a young start-up named Linkfluence,

which had just made a name for itself by announcing the victory of the no vote in the referendum on the TCE, by mapping the online debate within the French blogosphere.

To this day, Linkfluence plays a key role on the French and international market. It emerged from a group of students from the Technological University of Compiègne, led by computer and communication researcher Franck Ghitalla, and inspired by early works in the area of link studies (Adamic and Glance, 2005; Kleinberg, 1999). From January 2005, the little group had carried out a cartographic study of online debates about the TCE. They identified early the edge that supporters of the no vote had within the blogosphere, which contrasted greatly with the mainstream French media and early polls, clearly oriented towards the yes (Fouetillou, 2008). This precocious insight on what turned out to be a massive victory for the no vote gave these students an appreciable media coverage, including *Le Monde* and *The Financial Times*. The interest of the press in their method persuaded them of a potential commercial success: thanks to their new media connections, UTC students met Benoit Thieulin, who had them hired in 2006 by the GIS to pursue – in a more equipped and methodical way – the work he started two years earlier.

The next year, Linkfluence worked for the digital campaigning of presidential candidate Ségolène Royal, thanks to Thieulin's network again; meanwhile, they launched a blog called *The Presidential Observatory*, displaying daily each candidate's share of voice in the political blogosphere. They were not the only ones: the presidential election of 2007 was in fact a springboard for various experiments in this area. Scanblog, an agency founded in 2005 by two ex-marketers, launched in 2007 a blog called Politibuzz, which also ranked the most discussed candidates in French blogs. As for traditional pollsters, however, this public showcase was only a part of their work on the presidential election, since Scanblog also worked for one of the candidates. Net Conversations was another similar case, founded after the election by an ex-advisor of candidate François Bayrou, which had previously imagined an apparatus for monitoring opinions expressed in blogs about the centrist candidate:

> We had zero research budget. We didn't buy, or almost didn't buy polls and stuff. When you have such a small budget, what you do is listen to the web. So, I developed […] methods to monitor what was said on the Internet. I was following, I think, 500 sources, about 500 sources, there wasn't even a crawler, there was an RSS feed that worked, and most of the sources were blogs – it was the blog fever. So an RSS feed aggregator, we received every morning, and we read. – Founder of Net Conversations (interview, July 2014).

The presidential election provided emergent initiatives in online opinion monitoring a favorable context to get some funds and publicity, and to experiment their methodologies. After the election, having had their various achievements within the political sphere, Scanblog, Linkfluence and the others naturally moved towards corporate clients.

Sampling influencers: opinion measurement towards digital PR

They then met a powerful social demand that shaped their tools in a decisive manner: the growing concern of large companies for their online reputation (Power et al., 2009). In the mid-2000s, many marketing and communication professionals saw the explosion of online expression as a risk for corporate image, therefore exposed to all kinds of public criticism (Chun and Davies, 2001; McDonald and Slawson, 2002). In France, this topic was massively invested, especially after a momentous blog post by marketing specialist Christophe Deschamps in 2006, dubbing online reputation management as '*the next big thing*' for brands (4). Following this tendency, an abundant professional literature at the time described how the social web made consumers the

judges of corporate actions, whose verdicts would then compel and destabilize even the most dominant firms (Heiderich, 2004, 2009).

In this context, identifying opinion leaders able to shape brand's reputations, and report on their endeavors, was a crucial task. It became the first job of these online opinion pioneers. However, influence-based sampling also came from the scarcity of their technological resources. Founders of these young firms were indeed more skilled in marketing and communication than in computer programming. They therefore created consulting agencies, essentially based on their human workforce for analysis and consulting, equipped with free or low cost tools such as SocialMention, TweetDeck and RSS feed aggregators, essentially for data acquisition. Consequently, their business model demanded to limit the amounts of data their analysts had to deal with. In a context of a 'data deluge', and with limited tools to exploit them, their economical survival required them to build rules and models to sample the most valuable data.

Clients' demands and material constraints thus converged towards the sampling of the most influential people on a given subject, to collect and analyze their discourses. It was both a material necessity and a judicious idea: according to the persons I interviewed, not all opinions were equally interesting on the Internet. Some blogs, forums and websites were very popular, well-linked and easily accessible, while most of them remained unnoticed (Cardon, 2008). '*Listening to everyone*', in this perspective, would make the client '*hear nothing*', as Linkfluence's tagline of the time said. It would have mixed highly and poorly visible opinions together, blurring the measurement of reputation. This approach was then translated into the design of technology. Scanblog for example thus built a simple algorithm based on what they deemed to be the main factor of visibility and value for an online source: its position within search results.

> Our method is very silly, it is based on the assumption that the French Internet user still goes through search engines to get information. So on the one hand, it allows us to find out what people are saying about the brand, so it allows us to study their opinion. And, as we go through search engines, which is what the average Internet user does, we also see which material he is mainly exposed to, which allows us to measure reputation. – Co-founder of Scanblog (Interview, August 2012).

Along with early digital marketing publications (August, 2004; Kozinets, 2002), Scanblog sought to get insights on the reputation of the entities being discussed, by focusing on the first results of a search query. They built a heuristic approach that reproduced the average Internet user's behavior, to get a sense of the client's image circulating in the most visible publications and conversations. Such an approach was also used by Linkfluence, which built an "influence score", admittedly inspired by Google's PageRank (Brin and Page, 1998), designed to rank the most important protagonists within online informational ecosystems.

> We had no ambition of exhaustiveness; we wanted to work only on a sample of the websites that shaped opinions. So, we had to do a selection work. We worked with our crawling method to build a dynamic, constantly changing sample of the web, the 15,000 most influential sites on the French web – we gradually reached 15,000, it stabilized. And these 15,000 sites make the opinion of the French web. […] Our tool allows understanding where opinion is created, where the image is built. If I add a source that no one reads, it will distort the results. – Co-founder of Linkfluence (interview, December 2012).

Sampling the most visible sources was, in their perspective, inevitable, so much that it became a key epistemic feature of the model developed by Linkfluence. It was the way to select, in the chaos of online discourses, only the most valuable ones, those able to influence their audiences, prescribing attitudes and thus participating in the construction of the 'image' of a personality, a brand or a product. In this perspective, the unequally visible opinions produced by people interact and agglomerate to produce online reputations. In this perspective, selecting the most influential ones was then a necessary and sufficient condition for the elaboration of a good measure.

Reputation management as real-time exhaustive monitoring of online opinions

At the same time, a second (rival) approach to social listening emerged, led by technological start-ups whose founders were trained in computer sciences, and tried to develop automatic tools for parsing the web, indexing and analyzing the largest possible volumes of textual data produced by internet users. They all adopted a "software as a service" business model, soon to be paradigmatic in the software industry (Campbell-Kelly and Garcia-Swartz, 2007). Although they took a quite different approach from their sampling competitors, the software they elaborated were also significantly shaped by the demands of early clients for online reputation management tools. This second trajectory produced yet another way of measuring online opinions: software based on extensive databases and real-time dashboards, used for the early detection of online criticism and potential reputation crises.

Clients shaping the market: the influence of news monitoring

In the second half of the 2000s, numerous start-ups emerged, in France, USA, Canada and the United Kingdom, whose founders came from the world of software engineering. Unlike the agencies described in the previous section, these pioneers did not come from politics and classical public opinion. They were fascinated by the possibility of indexing and analyzing large datasets of public discourses on blogs and forums, and convinced that it could be of interest to many different kinds of customers. Indeed, as the sociology of innovation frequently shows (Akrich et al., 2006; Callon, 1979), their tools went through a lot of different uses before stabilizing. Sysomos emerged from a computer science project at University of Toronto, called BlogScope, aimed at indexing and analyzing the blogosphere. In a conference paper, its founders wrote: '*collecting, monitoring and analyzing information on blogs can provide key insights on "public opinion" on a variety of topics, such as products, political views, or entertainment*' (Bansal and Koudas, 2007: 1). However, other start-ups created in the same period were much less interested into public opinion measurement *per se*. Early leaders of the social media monitoring market Radian6 initially planned to sell their software to the police and counter-terrorism for surveillance; Brandwatch started working for the British government in 2005. However, all these companies would find their success in reputation monitoring.

In the case of the French-based global leader Synthesio, the software was born from an information aggregator built by its co-founder Thibault Hanin's, for his personal use at first, using crawlers to automatically retrieve and index news articles from online media. Months later, while getting a MBA at ESSEC business school after his engineer training, Thibault Hanin met Loïc Moisand, and started developing this idea with him, at first as a final-year project of "junior enterprise", then as a start-up of its own. '*That's when we explored the most options, clearly, between 2006 and 2008, we really tried a billion different things with this technology. Even today, if you wanted to do different things with it, you could*' (Hanin, interview, March 2015). Moisand and Hanin considered creating customized portals, popularity barometers, news aggregators. The easiness of indexing thousands of media, blog and forum posts allowed the two students to build indexing tools, graphs, charts and scores, although it was not yet clear what to do with it. The decisive encounter was the moment Synthesio signed an exclusivity contract with Kantar Media, a major player within the press clipping and audience measurement industry. Historically in charge of monitoring a company's public image in the media, Kantar found itself quite unprepared to deal with the explosion of online expression in the mid-2000s:

> Our first real product, finally, was simply a monitoring platform. Where the brands could put their subjects, and then it brought them everything. So, it was a lot of PR, a lot of communication… It was something that every big company in the world had on their desks, those press clipper things, but

online. At the time, there was also *L'Argus de la Presse* (5)… We worked with Kantar Media. They loved our thing, and they told us: we don't know how to manage the social web, you are the experts on that. So, we worked together, and that's how it took off. – T. Hanin, interview (March 2015).

The technology Synthesio developed to crawl media, blogs and forums allowed the extension of Kantar's traditional services of media monitoring to the social web of blogs and forums, and later on, social media. During almost two years, Kantar could then quantify its clients' media coverage, both online and offline. This partnership also provided a guaranteed source of revenue for the young start-up, thanks to the large clientele of their much larger partner. The first use of Synthesio's dashboards and databases was about providing large companies with tools to break new ground in a context where, increasingly, everyone became her own media and had to be treated as such.

'Listening to everything, everywhere, all the time' (6)*: SML as real-time exhaustive monitoring*

Originally used as a mere extension of press clipping services, Synthesio quickly caught the attention of the corporate communication departments of their clients that, as we saw previously, were extremely preoccupied by consumer criticism on the internet. Used at first to count the occurrences of a brand or product on the internet and measure media coverage, Synthesios's software became tools for monitoring the social web and detecting as soon as possible internet users' criticism against a company, its chief executives or its products, that may threaten the brand image. As the partnership with Kantar ended in early 2009, Synthesio started to sell its software directly to communication and PR departments.

> It came at some point from the customers, too. It's always the same story: our stuff was displayed everywhere, the communication guys saw it and said: ok wait, it gives me an idea… It was serendipity! Our dashboards were public at the time; there was a public URL that we could share. In terms of security, we could no longer do that today, but it circulated from one service to another, and people said: that's great, I too can use this information. – T. Hanin, interview (March 2015).

Meanwhile, the emerging SML software industry had been reinforced by the arrival of older companies, born in the 1990s and early 2000s: competitive intelligence and news monitoring software vendors, such as Meltwater (Norway, USA) or Digimind (France). These companies used to provide complex enterprise software for monitoring public or fee-based information sources, such as news broadcasts, patent databases, corporate and financial information, or scientific journals. From 2007 onwards, the attractive growth rates displayed by start-ups like Synthesio, and the pressing request of their clients, strove these companies to adapt their software to go into social data. Driven by the collective turmoil around online reputation, competitive intelligence software vendors redirected their activities, using their already-built technical infrastructures (servers, algorithms for processing and visualizing large amounts of data, text mining tools) to build up social media monitoring software for brand reputation management. By catching on the dynamics around social data, these vendors, with their established sales network and software infrastructures, contributed to stabilize the emerging real-time large-scale monitoring format for SML software.

> We used to do online monitoring, which meant competitive, product and technological monitoring. It was classic intelligence stuff. Then with online reputation we focused more on visibility, brand communication. We got into this thanks to Microsoft: they contacted us in early 2007, they wanted a software platform to monitor what was being said about them in France, both about their brand and their products. – Senior consultant at Digimind (interview, May 2012).

The technological skills of pioneering start-ups, and the arrival of competitive intelligence software on the market, thus resulted in the emergence of a specific standard: dashboards for

online reputation monitoring, based on the extensive and continuous indexation of data from blogs, forums and social platforms. As T. Hanin recalls, '*It was the fear of that time, not to miss anything within the data volumes*' (interview, March 2015). The first key task developers at Synthesio undertook was the construction of crawlers capable of indexing every possible source of social data.

> Our approach was to cover everything, everything that moves. So we started with the mainstream media, every news site, press, radio, we had a crazy categorization work there […]. Then we did blogs, of course; it was easy. Forums: much harder. It took us until mid-2008. – T. Hanin, interview (March 2015).

Figure 1. Example of a Synthesio dashboard (source: Synthesio's public communication).

The much larger volumes of data processed by SML editors thus oriented innovation towards the development of dashboard interfaces (see Figure 1), which included a number of analytic tools meant to allow the modelling and manipulation of data by the user, whom inherited the responsibility of making sense of them. As with other big data instruments, the preferred mode of inquiry in these tools became the search for correlations between different variables: temporality, semantics, type of sources, etc. (Dagiral and Parasie, 2017; Rieder, 2012). The user is offered scores and calculators that allow her to quantify the data associated with a particular keyword or search query over time, or the 'burst' of unexpected topics within online publications (Bansal and Koudas, 2007). To interface the user with the database, by giving her a series of cognitive holds on the abundant data, the dashboard format soon included to a similar set of analytical tools: charts, graphs, keyword clouds, verbatim threads, lists of authors, influence rankings and sentiment analysis scores being the most notable.

A hybrid quantification of online opinion and reputation

The contingent innovative trajectories followed by the pioneers of online opinion measurement are rooted in heterogeneous resources and constraints, and they produced different models. Trained in marketing and communication, evolving within the spheres of political counseling, tenants of the first approach developed a consulting model, heavily dependent on human workforce. Sampling online opinions seemed to them an efficient and judicious way to deal with the amounts of data produced in our daily conversations on the social web. On the other hand, the

computer scientists of the second approach naturally turned to the programming of automated software, able to retrieve and aggregate massive amounts of data, clearly belonging to the worlds of big data analytics (boyd and Crawford, 2012; Kitchin, 2014). These two models were competing for several years, until SML software gained decisive edge thanks to their much superior scalability, and to the massive financing they received as a result from tech venture capitalists (Author, 2018). Both models nevertheless converge in a significant manner, that I will discuss in this section.

The first similarity between them is how they were both shaped by their early clients. At a time when the explosion of online expression was perceived by many communication professionals as a destabilization to their methods, if not as a threat, social listening start-ups and agencies managed to leverage this uncertainty to sell their expertise, enacting what Boczkowksi (2004) calls, in the case of newspapers, a 'reactive, defensive and pragmatic' tackle of digitization. Beyond their diversity, both trajectories represent an example of how big data instruments are shaped by economical, cultural and material constraints, and point to the crucial role played by clients much larger than their service providers in the definition of what is relevant in an innovation, and what is not (Van Couvering, 2007). Combined with their epistemic and material resources, this injunction to measure and preserve the client's reputation on the Internet led agencies and software editors to produce models that both hybridize two previously separate epistemic traditions: that of public opinion assessment, and that of public relations.

As shown by historical studies, modern opinion polls were born in the 1930s at the crossroads of journalism, quantitative social psychology and statistics (Blondiaux, 1998; Herbst, 1995). They drew on the then recent findings discussed within the International Statistical Institute, then formalized by a canonical article from Jerzy Neyman, which posited that a key condition to build a representative sample of a given population was its randomness, i.e. that every individual within this population should have an equal probability of being selected (Neyman, 1934). This mathematical criteria is what allowed then George Gallup to describe his samples as a 'miniaturized America', in which everyone had a chance to participate, and every participant had the same weight (one person, one vote). This mimicry of the electoral process, and the powerfully democratic justification it brought to this new method, were essential to the success of opinion polls, and to their diffusion within the worlds of politics, media, but also academia and marketing (Blondiaux, 1998).

This mathematic and democratic cornerstone of opinion assessment is wiped away by the new methods developed by social listening agencies and software. On the one hand, influence-based sampling relies on an alternate definition of representativeness, which changes its political substance (Didier, 2013). Unlike traditional representative sampling, influence-based sampling implies a dissymmetric modeling of opinion, where only a minority of 'influencers' is given a prominent position in the mediated process of opinion construction. By adopting this hierarchical conception of opinion, sampling agencies embedded in their instruments the well-known power law distribution of attention on the internet (Kleinberg, 1999): only a small minority of sources are highly connected, visible and read. It also was part, at that time, of a larger movement of rediscovering of Lazarsfeld's 'two-step flow' of communication (Katz and Lazarsfeld, 2005) within the worlds of digital marketing (Mellet, 2009). It went hand in hand with the emergence at the same time of and influencer marketing, which extended the devoted care that once was the privilege of professional journalists, to famous bloggers (Charest and Bouffard, 2015).

Accordingly, SML actors' ambition of exhaustiveness and real-time indexing of online opinions, essential to the defensive monitoring perspective that shaped their software, also redefines the

way opinions are selected and analyzed by their users. By default, these instruments value the most visible, numerous and frequent opinions, and even more so, opinions that are becoming so, that are generating 'bursts' in the monitoring graphs and indicators. In the constant flow of online publications, SML software excel at detecting variations, ruptures, i.e. opinions that propagate and mobilize growing masses of Internet users. In this respect, they work in a similar way to Twitter's 'trending topics' algorithm, which favors the speed of diffusion of a hashtag more than its total volume (Gillespie, 2011). In this culture of 'attentional peaks' (Cardon, 2015: 83), attention is focused on what Boullier (2017) calls 'replications' (or 'aftershocks') of online social activity: the momentary aggregations of the conversation around a particular object. By focusing on the newest, fastest-diffusing opinions, those who – again – attract the most visibility and word-of-mouth, SML dashboards deliberately overlook constant trends and low-intensity opinions, expressed without any particular dynamic of expansion. The proverbial silence of the majorities then condemns them to disappear from the dashboards of social listening software.

Although they come from different perspectives, both models radically break with the statistical concept of representativeness that is the cornerstone of opinion polls, in order to take advantage of an abundant and 'post-demographic' conversational material (boyd and Crawford, 2012; Rogers, 2009). Agencies and software vendors share a dissymmetric conception of opinion measurement, where not all people and publications are of equal interest: in the sample-based approach, only the most influential ones are selected and analyzed; for their part, social listening platforms continuously seek the early detection of criticisms, to detect potential reputational crises. Both embed the mediated context of opinion production on the Internet within the quantification device. Opinion is not measured as an individual state of mind, but as a collective and mediated activity, taking into account the online ecological contexts of its production, and the unequal power of mobilization of online publics.

Social media listening thus hybridizes opinion measurement with a public relations paradigm, by weighing the publicly stated expressions of ordinary people on the internet according to their supposed reach and influence. It values the most visible and viral discourses, which are most likely to weigh in the construction of reputation. Unlike opinion polls, these instruments do not produce 'political rationality' (Blondiaux, 2004), which explains why they do not compete with them on matters of public opinion, and remain confined to strategic uses, behind the closed doors of communication and marketing departments. By modeling opinion through techniques and formats that have long been used in the management of large companies, such as dashboards and databases (Yates, 1993), these instruments enact a very different regime than that of public opinion, understood as the normative and legitimate reference of public debate. Closely linked to the problems and practical knowledge of marketing and communication services, SML thus tends to become an internal resource for large companies, and participates in the instrumentation of brand image management, considered as reputation capital (Power et al., 2009).

Conclusion

While professional pollsters largely ignored James R. Beniger's intuition, the opportunity of online opinion measurement was seized by agencies and startups that adopted two main trajectories in their innovation efforts. The first one shows some similarities with the traditional conception of opinion quantification: it started in the political sphere, and was concerned with the elaboration of conventions for representativeness – although in a renewed manner. Influence-based sampling is the way agencies with low technological capital overcame their technical limitations and defined the epistemic value of their contribution: the measurement of public image

within an elite of influential bloggers. On the other hand, the trajectory of social listening startups was based on the massive and deliberately indistinct capturing of all kinds of social data, regardless of their influence of visibility. It was a crucial matter for these actors to index a maximum of data to display within their dashboards, conceived as panoptic and synthetic tools for continuous reputation monitoring. As shown above, both approaches hybridize the measurement of public opinion with corporate concerns and knowledge devices, inscribed in the field of PR and reputation management.

It is difficult at this point to say how typical of the French market this process is. The 'big data' approach is now dominant at an international level (Author, 2018), but the sampled approach it had to compete with in France has not been identified in another country, notably in the USA where I conducted interviews in 2014. Although this handcrafted method was ultimately wiped off, its persistence throughout the years can be linked to the French historical valorization of columnists and opinion leaders (Karila-Cohen, 2008), also perceptible in the earliness and importance of the French political blogosphere (Siles, 2017). As with the case of blogging, it seems that the French technologies for online opinion monitoring aligned progressively with international – that is, mostly American – standards and values. This article then calls for symmetric investigation in the USA and other countries where major players of the social listening industry were born.

Ultimately, this investigation shows the importance of taking into account the multiple and parallel histories of big data, including even the history of 'small' or 'medium data' analytics, and semi-automatic attempts that once existed, before 'big data' and its particular set of standards became conventional, (Kitchin, 2014; Rieder, 2012; Schutt and O'Neil, 2013). In fact, the trajectories studied in this paper are intertwined, and borrow from each other: still today, social listening platforms conventionally include influence scoring and viral marketing management tools that clearly derive from the hierarchized, digital-PR, conception of online opinion. In this regard, this research also shows how ancient traditions such as PR and news clipping are inscribed within big data tools and how they shape their epistemology and their uses. This is a crucial point if we wish to understand the cultural and material conditions of reception of data science in the many different social worlds in which it is increasingly used (Dagiral and Parasie, 2017).

While there is a legitimate preoccupation about the "capitalism of surveillance" (Zuboff, 2018), one research priority should be to understand the genealogy and methodological premises of the device that materially perform said surveillance. Future research should also produce in-depth accounts of the actual uses of big data surveillance technologies, since they cannot be understood simply by studying their general working principles (Jaton, 2017; Seaver, 2017). Historical and ethnographic studies of algorithmic devices are key if we are to understand how data is constituted as a new form of capital by corporate actors.

Notes

(1) According to Markets & Markets, the social media analytics market generated $2.71 billion revenue in 2017, and is estimated to increase up to more than $9 billion in 2022. Source: https://www.marketsandmarkets.com/PressReleases/social-media-analytics.asp [consulted on 06/18/19].

(2) A search on the Web of Science (TS= "social media analytics" OR "social listening" OR "social media analysis") gathered only 132 articles in social sciences since 2007, none of them taking social listening as an object of study.

(3) See the latest report by Forrester Research, "The Forrester Wave™: Social Listening Platforms, Q3 2018".

(4) Text available (in French): https://www.zdnet.fr/blogs/ils-ont-blogue/l-indispensable-gestion-de-la-reputation-numerique-39600489.htm

(5) Historical player of French news monitoring and PR founded in 1879, L'Argus de la Presse was bought by software company Cision in 2017.

(6) Brandwatch's motto when they opened their first Parisian office in 2015.

REFERENCES

Adamic, L.A. and N. Glance (2005) 'The political blogosphere and the 2004 US election: divided they blog', *Proceedings of the 3rd international workshop on Link discovery*, New-York, USA, 2005, pp. 36–43.

Akrich, M. M. Callon and B. Latour (2006) *Sociologie de La Traduction. Textes Fondateurs*. Paris:Presses de l'École des Mines.

Anderson, C. (2008) 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired*, 23 June. Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

August, S. (2004) 'Community Involvment: Using Online Ethnography to Explore the Digital Camera Market', *Quirks Marketing Research Review*, July.

Bansal, N. and N. Koudas (2007) 'Searching the Blogosphere', *Proceedings of the 10th international Workshop on Web and Databases*, Beijing, China, 2007.

Beer, D. (2016a) 'How should we do the history of Big Data?', *Big Data & Society* 3(1): 1–10.

Beer, D. (2016b) 'The Data Analytics Industry and the Promises of Realtime Knowing: Perpetuating and Deploying a Rationality of Speed', *Journal of Cultural Economy* 10(1): 21-33.

Beniger, J.R. (1998) 'Presidential Address: Survey and Market Research Confront Their Futures on the World Wide Web', *Public Opinion Quarterly* 62(3): 442–452.

Blondiaux, L. (1998) *La Fabrique de l'opinion. Une Histoire Sociale Des Sondages*. Paris: Seuil.

Blondiaux, L. (2004) 'Le nouveau régime des opinions', *Mil Neuf Cent, Revue d'Histoire Intellectuelle* 22(1): 161–171.

Blumer, H. (1948) 'Public Opinion and Public Opinion Polling', *American Sociological Review* 13(5): 542–549.

Boczkowski, P.J. (2004) *Digitizing the News: Innovation in Online Newspapers*. Cambridge, Mass: MIT Press.

Boullier, D. (2017) 'Big data challenges for social sciences: from society and opinion to replications', *ISA e-symposium*, 2017.

Bourdieu, P. (1979) 'Public opinion does not exist', pp. 124–130 in A. Mattelart and S. Siegelaub (eds) *Communication and Class Struggle. Capitalism, Imperialism*. New York: International General.

boyd, d. and K. Crawford (2012) 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon', *Information, Communication & Society* 15(5): 662–679.

Brin, S. and L. Page (1998) 'The Anatomy of a Large-Scale Hypertextual Web Search Engine', *Proceedings of the seventh international conference on World Wide Web*, 1998.

Brooker, P., J. Barnett and T. Cribbin (2016) 'Doing social media analytics', *Big Data & Society* 3(2): 1–12.

Callon, M. (1979) 'L'État face à l'innovation technique : le cas du véhicule électrique', *Revue française de science politique* 29(3): 426–447.

Campbell-Kelly, M. and D. D. Garcia-Swartz (2007) 'From Products to Services: The Software Industry in the Internet Era', *Business History Review* 81(4): 735–764.

Cardon, D. (2008) 'Le design de la visibilité', *Réseaux* 152(6): 93–137.

Cardon, D. (2010) *La Démocratie Internet. Promesses et Limites*. Paris: Seuil.

Cardon, D. (2015) *À quoi Rêvent les Algorithmes. Nos Vies à l'Heure des Big Data*. Paris: Seuil.

Ceron, A., L. Curini and S.M. Iacus (2014) 'Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France', *New Media & Society* 16(2): 340–358.

Charest, F. and J. Bouffard (2015) 'The Characteristics of the e-Influence of Community Managers: Issues for the e-Reputation of Organizations', *Public Relations Review* 41(2): 302-304.

Chun, R. and G. Davies (2001) 'E-reputation: The role of mission and vision statements in positioning strategy', *The Journal of Brand Management* 8: 315–333.

Dagiral, É. and S. Parasie (2017) 'La « Science des Données » à la Conquête des Mondes Sociaux. Ce que le « Big Data » Doit aux Épistémologies Locales', pp. 85–104 in P.-M.

Menger and S. Paye (eds) *Big Data et Traçabilité Numérique. Les Sciences Sociales Face à La Quantification Massive Des Individus*. Paris: Collège de France.

Didier, E. (2013) 'Histoire de la représentativité statistique : quand le politique refait toujours surface', pp.15-30 in *La Représentativité En Statistique*. Paris: PUF/INED.

Flichy, P. (2007) *Understanding Technological Innovation. A Socio-Technical Approach*. Northampton, MA: Edward Elgar.

Fouetillou, G. (2008) 'Le web et le traité constitutionnel européen. Écologie d'une localité thématique compétitive', *Réseaux* 147(1): 229–257.

Gayo-Avello, D. (2013) 'A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data', *Social Science Computer Review* 31(6): 649–679.

Gillespie, T. (2011) 'Can an algorithm be wrong? Twitter Trends, the specter of censorship, and our faith in the algorithms around us', in *Culture Digitally*. Available at: http://culturedigitally.org/2011/10/can-an-algorithm-be-wrong/ (accessed 1 May 2020).

Gillespie, T. (2014) 'The relevance of algorithms', pp. 167–193 in P. Boczkowski, T. Gillespie, and K.A. Foot (eds) *Media Technologies. Essays on Communication, Materiality and Society*. Boston: MIT Press.

Heiderich, D. (2004) *Rumeur sur Internet: Comprendre, Anticiper et Gérer les Cybercrises*. Paris: Village mondial.

Heiderich, D. (2009) 'Influence sur Internet. Perceptions et mécanismes d'influence sur Internet dans la société de l'urgence'. *Magazine de la communication de crise et sensible*. Available at: http://www.communication-sensible.com/download/influence-sur-internet-didier-heiderich.pdf

Herbst, S. (1995) *Numbered Voices. How Opinion Polling Has Shaped American Politics*. Chicago: The University of Chicago Press.

Jaton, F. (2017) 'We get the algorithms of our ground truths. Designing referential databases in digital image processing', *Social Studies of Science* 47(6): 811–840.

Karila-Cohen, P. (2008) *L'État Des Esprits. L'Invention de l'Enquête Politique En France (1814-1848)*. Rennes: Presses Universitaires de Rennes.

Katz, E. and P. Lazarsfeld (2005) *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers.

Kitchin, R. (2014) 'Big Data, New Epistemologies and Paradigm Shifts', *Big Data & Society* 1(1): 1–12.

Kleinberg, J. (1999) 'Authoritative sources in a hyperlinked environnment', *Journal of the ACM* 46(5): 604–632.

Knorr-Cetina, K. (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kozinets, R.V. (2002) 'The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities', *Journal of Marketing Research* 39: 61–72.

Kreiss, D. (2012) *Taking Our Country Back: The Crafting of Networked Politics from Howard Dean to Barack Obama*. Oxford: Oxford University Press.

Latour, B. (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford ; New York: Oxford University Press.

Lyon, D. (ed.) (2003) *Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination*. London ; New York: Routledge.

MacKenzie, D. (2011) 'Evaluation Cultures? On Invoking 'Cultures' in the Analysis of Behaviour in Financial Markets', *Working paper*. Available at: http://www.sps.ed.ac.uk/__data/assets/pdf_file/0007/64564/EvalCults11.pdf (accessed 2 November 2017).

Mayer-Schönberger, V. and K. Cukier (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

McDonald C.G. and V.C. Slawson (2002) 'Reputation in an Internet Auction Market', *Economic Inquiry* 40(4): 633–650.

Mellet, K. (2009) 'Aux sources du marketing viral', *Réseaux* 157(5): 267–292.

Neyman, J. (1934) 'On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection', *Journal of the Royal Statistical Society* 97(4): 558–625.

Power, M., T. Scheytt and K. Soin (2009) 'Reputational Risk as a Logic of Organizing in Late Modernity', *Organization Studies* 30(2–3): 301–324.

Pridmore, J. and D. Zwick (2011) 'Marketing and the Rise of Commercial Consumer Surveillance', *Surveillance & Society* 8(3): 269–277.

Quéré, L. (1990) 'Opinion : l'Économie du Vraisemblable. Introduction à une Approche Praxéologique de l'Opinion Publique', *Réseaux* 43(8): 33–58.

Rieder, B. (2012) 'ORDER BY column_name. The Relational Database as Pervasive Cultural Form' in *The Lived Logics of Database Machinery* (workshop), London. Available at: http://thepoliticsofsystems.net/2012/06/the-lived-logics-of-database-machinery/

Rogers, R. (2009) 'Post-Demographic Machines', pp. 29–39 in Annet Dekker (ed) *Walled Garden*. Amsterdam: Virtueel Platform.

Rogers, R. and N. Marres (2000) 'Landscaping Climate Change: a Mapping Technique for Understanding Science and Technology Debates on the World Wide Web', *Public Understanding of Science* 9(2): 141–163.

Rouvroy, A. (2018) 'Governing Without Norms: Algorithmic Governmentality', *Psychoanalytical Notebooks* 32(1): 99-111.

Schutt, R. and C. O'Neil (2013) *Doing Data Science. Straight Talk From the Frontline*. Sebastopol: O'Reilly Media.

Seaver, N. (2017) 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems', *Big Data & Society* 4(2): 1–12.

Siles, I. (2017) *Networked Selves: Trajectories of Blogging in the United States and France*. New York: Peter Lang.

Trottier, D. (2012) *Social Media as Surveillance: Rethinking Visibility in a Converging World*. Farnham: Ashgate.

Van Couvering, E. (2007) 'Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality', *Journal of Computer-Mediated Communication* 12(3): 866–887.

Vargo, C.J., L. Guo and M.A. Amazeen (2018) The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society* 20(5): 2028–2049.

Watts, D.J. and P.S. Dodds (2007) 'Influentials, Networks, and Public Opinion Formation', *Journal of Consumer Research* 34: 441–457.

Whyte, C.E. (2016) 'Thinking inside the (black) box: Agenda setting, information seeking, and the marketplace of ideas in the 2012 presidential election', *New Media & Society* 18(8): 1680–1697.

Yates, J. (1993) *Control Through Communication: The Rise of System in American Management*. Baltimore, MD: Johns Hopkins University Press.

Yeung, K. (2018) 'Algorithmic Regulation: A critical interrogation' *Regulation & Governance* 12(4): 505–523.

Zuboff S (2018) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.