Regular Article

new media & society
2014, Vol. 16(2) 340–358

DOI: 10.1177/1461444813480466
nms.sagepub.com
SAGE

Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France

Andrea Ceron, Luigi Curini, Stefano M Iacus
Università degli Studi di Milano, Italy

Giuseppe Porro
Università degli Studi dell'Insubria, Italy

Abstract
The growing usage of social media by a wider audience of citizens sharply increases the possibility of investigating the web as a device to explore and track political preferences. In the present paper we apply a method recently proposed by other social scientists to three different scenarios, by analyzing on one side the online popularity of Italian political leaders throughout 2011, and on the other the voting intention of French Internet users in both the 2012 presidential ballot and the subsequent legislative election. While Internet users are not necessarily representative of the whole population of a country's citizens, our analysis shows a remarkable ability for social media to forecast electoral results, as well as a noteworthy correlation between social media and the results of traditional mass surveys. We also illustrate that the predictive ability of social media analysis strengthens as the number of citizens expressing their opinion online increases, provided that the citizens act consistently on these opinions.

Corresponding author:
Luigi Curini, Department of Social and Political Sciences, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milano, Italy.
Email: luigi.curini@unimi.it

Keywords
Analysis of public opinion, electoral campaign, political forecast, sentiment analysis, social media, text analytics, text mining

The exponential growth of social media and social networking sites such as Facebook and Twitter raises the possibility of using the web to explore and track the (political) preferences of citizens. The Internet indeed represents a valuable source of data that is useful for monitoring public opinion (Madge et al., 2009; Woodly, 2007). Due to recent developments in quantitative text analysis and "sentiment analysis" (SA), we are now better able to exploit such information in a reliable manner.

Scholars have recently begun to investigate the use of social media as a device to forecast elections (Tjong and Bos, 2012), assess the popularity of politicians (Gloor et al., 2009), and compare the political preferences that citizens express online with those captured using traditional polls (O'Connor et al., 2010). Some of these works rely on simple techniques, focusing on the volume of data related to specific parties or candidates. For instance, Véronis (2007) demonstrated that the number of candidate mentions in blog posts is a good predictor of electoral success and can perform better than election polls. Similarly, some scholars have claimed that a candidate's number of Facebook supporters could be a valid indicator of electoral fortunes (Upton, 2010; Williams and Gulati, 2008), while Tumasjan et al. (2010) compared party mentions on Twitter with the results of the 2009 German election and argued that the number of tweets related to each party is a good predictor of its vote share.

Nonetheless, not all enquiries have succeeded in correctly predicting the outcomes of elections (Gayo-Avello et al., 2011; Goldstein and Rainey, 2010). For instance, the share of campaign weblogs prior to the 2005 federal election in Germany was shown to be a poor predictor of the relative strength of the parties, insofar as small parties were over-represented (Albrecht et al., 2007). In a study on Canadian elections, Jansen and Koop (2005) failed in their estimation of the two largest parties' positions. Gayo-Avello (2011) demonstrated that social media analysts would have overestimated Obama's victory in 2008 (up to the point of predicting his success even in Texas). Jungherr et al. (2012) criticized the work of Tumasjan et al. (2010), arguing that including the German Pirate Party in the analysis would have had yielded a negative effect on the accuracy of the predictions.

It has also been noted that merely counting mentions or tweets is not sufficient to provide accurate foresight (Chung and Mustafaraj, 2011). Accordingly, other studies have attempted to improve this stream of research using SA.[1] Lindsay (2008), for example, developed a sentiment classifier based on lexical induction and found correlations between several polls conducted during the 2008 US presidential election and the content of wall posts available on Facebook. O'Connor et al. (2010) showed similar results, revealing a correlation between Obama's approval rate and the sentiment expressed by Twitter users. In addition, SA of tweets was found to perform as well as polls in predicting the results of the 2011 Dutch senate election (Sang and Bos, 2012), while analysis of multiple social media sites (Facebook, Twitter, Google, and YouTube) outperformed traditional surveys in estimating the results of the 2010 UK election (Franch, 2012).

In the present paper, we follow this latter strategy by adopting the method proposed in Hopkins and King (2010) (hereafter 'HK'). As we will discuss, this method presents

various advantages compared to traditional SA techniques. We will employ this method in three different scenarios, by tracking the online popularity of Italian political leaders throughout 2011 and the voting intentions of French Internet users in both the 2012 presidential election and in the subsequent legislative election. In all cases, we will contrast our results with those obtained through traditional offline surveys, as well as with actual electoral results. The selection of a variety of contexts to be analyzed here was intentional in order to better investigate the strengths and limits of monitoring social media, as well as to assess which factors can increase (or decrease) their reliability. In the conclusion, we propose some suggestions for future research.

How to scrutinize citizens' preferences through social media

Nowadays, Internet access is available to a wide audience of citizens; accordingly, social media usage is growing at a fast rate. Throughout the world, approximately 35 of every 100 people had access to the web in 2011 (approximately 2.5 billion people).[2] In total, 72% of the Internet population is active on at least one social network,[3] such as Facebook (over 800 million of users, 12% of the world's population)[4] or Twitter (140 million active users).[5]

Recently, social network sites have also begun to wield substantive effects on real-world politics: they have been used to organize demonstrations and revolts, such as during the 'Arab spring' (Cottle, 2011; Ghannam, 2011);[6] to engage individuals in mobilization (Bennett and Segerberg, 2011; Segerberg and Bennett, 2011); and to build social movements and political parties, such as the Pirate Party in Sweden and Germany or the Italian *Movimento 5 Stelle*, which uses the web to set the party line and to select candidates.[7]

Accordingly, there has been discussion about whether the web may become an "uncoerced public sphere" (Benkler, 2006; Downey and Fenton, 2003; Langman, 2005). While some authors suggest that the Internet and social media are potential sources of direct democracy that may contribute to increased responsiveness and accountability in real-world politics (De Zúñiga et al., 2009; Papacharissi, 2002), others have proposed diverging views strongly criticizing this same idea (Alvarez and Hall, 2011; Hindman, 2009; Larsson and Moe, 2012).

Notwithstanding this debate, given the large amount of data related to public opinion available online (and its growing relevance), monitoring this flow of preferences becomes an important task *per se*.[8] The challenge is to select the methods that are most appropriate in this regard. While earlier studies, as already discussed, focused mainly on the volume of data (related, for instance, to each party or candidate), here we aim to capture Internet users' attitudes in greater detail, beyond merely tabulating numbers of mentions. Accordingly, we will employ the method recently proposed by Hopkins and King (2010).

The main advantage of the HK method is that it performs a supervised SA. The traditional approach to SA is based on the use of ontological dictionaries: in other words, a text is assigned to a specific opinion category if some pre-determined words or expressions appear (or do not) in the text. The advantage of this approach is, of course, the possibility to implement a totally automated analysis (once the dictionary has been defined). However, a major drawback is the difficulty in classifying opinions expressed

through ironic or paradoxical sentences, or in appreciating all of the nuances of language (e.g., specific jargons, neologisms): the informal expression "what a nice rip-off!", for instance, is quite ambiguous from the viewpoint of an ontological dictionary, because it includes both a positive *and* a negative term.

On the contrary, the HK method relies on a two-stage process. The first step involves human coders and consists of reading and coding a subsample of the documents downloaded from some Internet source. This subsample—with no particular statistical property: see below—represents a training set that will be used by the HK algorithm to classify all the unread documents during the second stage. Human coders are, of course, more effective and careful than ontological dictionaries when it comes to recognizing all of the previously discussed language specificities and an author's attitude toward their subject. Moreover, human coding is better suited to identify the (ever-present) problem of spamming in social communication. This issue is of course important, given that spamming can impact the accuracy of the final result. At the second stage, the automated statistical analysis provided by the HK algorithm extends such accuracy to the entire population of posts, allowing for proper capture of the opinions expressed on the web. Indeed, the expected error of the estimate is approximately 3%.[9]

The methodology is based on the assumption that the opinions of people posting on social networks can be deduced by all of the terms they use—not only the terms explicitly related to the topic being discussed, but also the "neutral" part of the language that is commonly used. Therefore, to characterize different opinions, the single units (blogs, posts) in the data set are decomposed into their own single words and, consequently, each unit is represented by the vector of the terms used, which we call the "word profile" of the unit.[10]

The formal background of the method is simple (for further details see Hopkins and King, 2010). The word profiles used in the text units are indicated by **S**, and the opinions expressed by people posting the texts are indicated by **D**. The frequency distribution of the terms $P(\mathbf{S})$ can be expressed as:

$$(*) \qquad P(\mathbf{S}) = P(\mathbf{S}/\mathbf{D})\, P(\mathbf{D})$$

where $P(\mathbf{D})$ is the frequency distribution of the opinions

The aim of the method is to obtain an estimate of $P(\mathbf{D})$—that is, to know how the opinion is distributed over the posting population. The frequency distribution $P(\mathbf{S})$ can be evaluated by tabulating all the texts posted, requiring only some computer time and no debatable assumptions. The conditional distribution $P(\mathbf{S}|\mathbf{D})$ cannot be observed; it must be estimated by hand-coding the training set of texts.

The hand-coding of the training text allows for the calculation of $P_T(\mathbf{S}|\mathbf{D})$—that is, the conditional frequency distribution of word profiles inside the training set. The reasonable requirement of the method is that the texts of the training set are homogeneous across the whole data set; that is, they come from the same "world" as the rest of the dataset. If this is the case, the frequency distribution of the opinions can be consistently estimated, because both $P(\mathbf{S})$ and $P_T(\mathbf{S}|\mathbf{D})$ are observable. Therefore, by equation (*) and noticing that $P_T(\mathbf{S}|\mathbf{D})$ and $P(\mathbf{S}|\mathbf{D})$ are both matrixes, we have

$$P(\mathbf{D}) = P(\mathbf{S}|\mathbf{D})^{-1}\, P(\mathbf{S}) = P_T(\mathbf{S}|\mathbf{D})^{-1}\, P(\mathbf{S})$$

where $P_T(\mathbf{S}|\mathbf{D})^{-1}$ is the inverse matrix of $P_T(\mathbf{S}|\mathbf{D})$, and similarly for $P(\mathbf{S}|\mathbf{D})^{-1}$.

It is worth remarking that while homogeneity between the training set and the dataset is required, no statistical property must be satisfied by the set: in particular, the training set is not a representative sample of the population of texts.

Broadly speaking, there are several social media sites that can be analyzed. Here, we will focus on Twitter, a social network for microblogging (Jansen et al., 2009) that has experienced sharp growth in recent months. Today, Twitter is the second highest ranking social network, behind Facebook; in 2009, it ranked 22nd. With regard to the countries analyzed in this work, we observe that in June 2011, Twitter was the second highest ranking social networking site in France and the third most used social networking site in Italy. In particular, in February–March 2012, 12 million Italian users were active on Twitter (Mazzoleni et al., 2011). A further crucial advantage of Twitter, which gives it great popularity in the literature on social media analysis, is that most posts by users ("tweets" in Twitter jargon) are freely accessible, contrary to other social networks.

To download the data employed in the present paper, we relied on two sources: the social media monitoring and analytics platform *Crimson Hexagon* (http://www.crimsonhexagon.com/) and the Internet engine *Voices from the Blogs* (http://www.voicesfromtheblogs.com/).[11] For the Italian case, analysis of the tweets was performed directly using the ForSight platform provided by *Crimson Hexagon*, while for the two French cases, the data were collected through *Voices from the Blogs* and analyses were run in R (http://www.r-project.org/).[12]

Compared to traditional survey polls, running an analysis on social media is attractive for a number of reasons (Xin et al., 2010). First, social media analysis is cheaper and faster compared to traditional surveys, and enables continuous monitoring of public opinion by performing real-time analysis. On the contrary, offline surveys are—by definition—more static. This feature is particularly relevant during electoral campaigns, as we will discuss below. In fact, using SA, we can measure voters' attitude on a day-by-day basis. Hence, we are able to capture the reaction of public opinion to any exogenous stimulus by observing the shift in preferences measured immediately after the shock. Similarly, analyzing social media also allows us to observe trends and breaking points. This feature has obvious implications for both researchers and spin doctors. Scholars can benefit from the amount of information available to investigate preferences in the making, while analysts and advisors can exploit these data to adjust the frame of their electoral campaign.

In addition, traditional surveys pose solicited questions, and it is well known that this approach might inflate the share of strategic answers (Payne, 1951). Conversely, SA does not utilize questionnaires and focuses only on listening to the stream of unsolicited opinions freely expressed on the Internet. In other words, SA adopts a bottom-up approach, at least if compared with the more traditional top-down approach of offline surveys. Far from saying that all of the comments posted on social networks contain the sincere preferences of the author, we can argue that the Internet may represent, to a large extent, an arena in which users are free to express themselves (Savigny, 2002).[13] Thus, the social network should be in a position to be less affected by the spiral of silence (Noelle-Neumann, 1974).[14] Moreover, while web analysis must contend with the problem of silent users, surveys face the problem of low response rates.

The main weakness that is usually noted when talking about social media analysis is related to the question of whether users of social media are representative of the entire population. Indeed, although the number of social media users is rapidly growing, the socio-economic traits of citizens who have access to the web do not exactly match the actual demographics of the whole population (Sang and Bos, 2012). For example, previous studies show that senior citizens are under-represented on the web (Fox, 2010), and there is a prevalence of highly educated male individuals (Wei and Hindman, 2011).[15] Interestingly, the social-demographic differences are reduced when considering only the sample of people who express political opinions online (on this point see Bakker and De Vreese, 2011).[16] Similarly, Best and Krueger (2005: 204, italics added) underline that "although online participators currently skew in a liberal direction, the online environment, at least compared to the offline environment, only marginally advantages the political voice of *liberals*."

In summary, although the population of social media users is not representative of one country's citizenry, there are still some doubts about whether such bias could affect the *predictive skills* of social media analysis compared to traditional offline surveys. Indeed, the former aspect (the predictive skills of social-media analysis) does not necessarily require the previous factor (the issue of representation) to hold true in order to effectively apply. This scenario can happen, for example, if we assume that politically active Internet users act like opinion-makers who are able to influence (or to "anticipate") the preferences of a wider audience: consequently, it could be found that the preferences expressed through social media today will affect (predict) the opinion of the entire population tomorrow (O'Connor et al., 2010).

In the following sections, we will test the predictive skills of social media analysis by employing the HK method in two different countries (Italy and France) and over three distinct political phenomena: leaders' popularity and presidential and legislative national elections.

Comparing Italian leaders' popularity ratings in 2011

The first political context in which we explore the usefulness (and reliability) of social media analysis concerns the relationship between the popularity ratings of the main Italian political leaders throughout 2011 as determined by traditional mass surveys (source: ISPO, Istituto per gli Studi sulla Opinione Pubblica) and by analysis of social media posts. In this sense, we treat the former surveys as our benchmark, and control how closely the latter approach them.

The mass surveys' popularity ratings span 13 January 2011 to 20 October 2011. They are based on a sample of approximately 800 respondents, and focus on seven leaders: Silvio Berlusconi (the leader of the PDL and the Italian Prime Minister at that time), Pier Luigi Bersani (the leader of the PD, the main opposition party), Umberto Bossi (the leader of the Northern League and the main cabinet partner of the PDL at that time), Pier Ferdinando Casini (the leader of the centrist UDC, an opposition party), Antonio Di Pietro (the leader of the IDV, an opposition party), Gianfranco Fini (the President of the Italian Lower Chamber and co-founder of the PDL before leaving the party at the end of 2011), and Nichi Vendola (the leader of the radical-left party SEL, the main extra-parliamentary opposition party). The popularity ratings range from 0 to 100 and identify the percentage of positive scores given by the respondents to each leader.[17]

Similarly, the popularity of each leader according to social media has been estimated as the percentage of his positive posts over the sum of his positive and negative posts; this value once again ranges from 0 to 100 to make it comparable to the survey popularity ratings. We considered two different temporal ranges: in the first case, we collected all of the posts concerning each leader in the month preceding the day on which the mass survey was actually administered; in the second, we re-ran the above procedure considering just the week preceding the day on which the mass survey was administered. Overall, we analyzed over 107,000 tweets in the monthly timing and 32,000 tweets in the weekly timing. Given that the results of our analysis look remarkably similar regardless of the time period considered, we focus here on the popularity of scores that arose in the weekly timing (according to the choice made by Sang and Bos, 2012).

In Table 1 below, we report the average difference (mass surveys minus social media popularity ratings) of the scores so obtained. Three main findings clearly arise. First, provided that we consider all leaders without any internal distinction, the average mass survey ratings appear always to be higher than the social media ratings (by more than five points on average). This phenomenon is true for all of the leaders except Di Pietro and Fini, whose online popularity appears to be higher. Second, we find a considerable variation among the leaders: for example, the average difference between the two measures of ratings is quite low for both Bossi and Fini (albeit with a different sign in the two cases), while it increases considerably for Casini, Bersani, and Vendola. Third, the correlation between the mass surveys and social media ratings is positive, albeit not dramatically strong. Note, however, that there is a marked contrast between Berlusconi, Bersani, and Bossi (the three most important and visible Italian leaders during 2011) and the remaining leaders. For our first set of leaders, the correlation is indeed considerably higher, particularly for Berlusconi ($r = .93$) and Bersani ($r = .75$).

Table 1. Average difference and correlation of leaders' popularity ratings between mass surveys and social media. Here *n* is the number of mass surveys for each leader.

	Average difference	*Standard deviation*	*R*	*n*
All leaders				
- % positive posts (previous week)	5.71	.497	.241	43
Berlusconi				
- % positive posts (previous week)	8.71	1.89	**.933**	7
Bersani				
- % positive posts (previous week)	12.53	1.40	**.746**	6
Bossi				
- % positive posts (previous week)	2.58	2.54	**.540**	6
Casini				
- % positive posts (previous week)	13.34	2.92	–.008	5
Di Pietro				
- % positive posts (previous week)	–4.12	1.97	.109	6
Fini				
- % positive posts (previous week)	–2.30	2.93	.005	7
Vendola				
- % positive posts (previous week)	10.32	2.14	.090	6

However, Table 1 only provides an aggregate (and static) picture that summarizes all of the temporal observations. Therefore, the table cannot tell us anything related to the *dynamic* relationship between our two measures of popularity ratings. To explore this issue, Figure 1 below plots the evolution over time of the Mean Absolute Error (MAE) of predictions of the leaders' popularity as they arise from social media as compared to the scores obtained from mass surveys. MAE has been widely used to compare the accuracy of forecasts based on social network analysis (Tumasjan et al., 2010). We also adopt this approach in this work. As can be seen, despite being quite relevant at the beginning of 2011 (approximately 13 points), the absolute difference tends to decrease markedly as time goes by.

In summary, our results provide two interesting insights:

- At least for the most visible leaders, the two measures of popularity ratings (mass surveys and social media) seem to be consistent; that is, they appear to react in the same way to exogenous factors (i.e., news reported in the media concerning particular leaders, political events, etc.). When the first measure increases, the second measure also increases.
- Mass surveys are on average more "generous" than social media with respect to popularity ratings (i.e., they generally give a higher rating to political leaders). However, the average difference between the two measures of popularity ratings, at least during 2011, seems to be clearly declining over time.

In this last respect, it is worth noting that the possibility of new elections was widely debated (and was a real possibility) during the second part of 2011, during the Italian

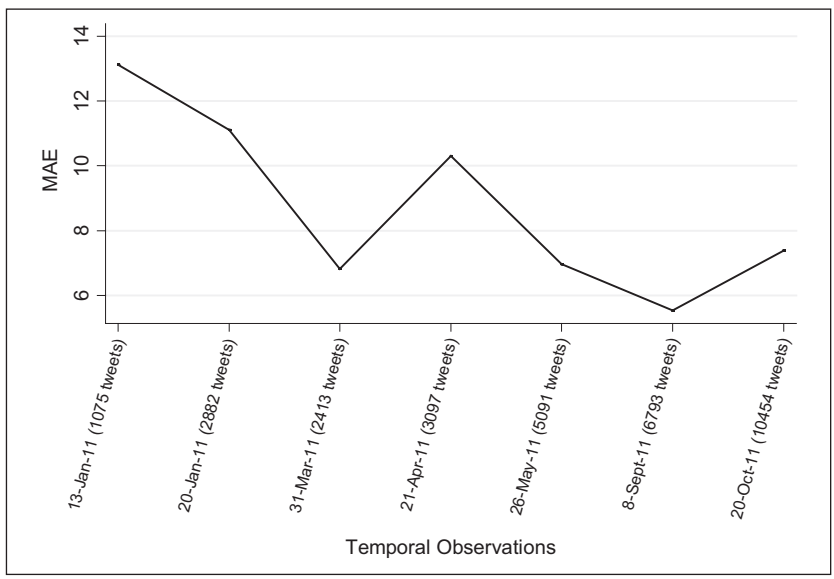

Figure 1. The MAE of leaders' popularity ratings between mass surveys and social media estimated over time.

political debates that took place throughout the Berlusconi IV cabinet's political crisis.[18] In this sense, it could be argued that as the shadow of an election approaches, more people tend to express their opinions on politics within social media (and indeed the number of tweets about Italian political leaders more than doubled, on average, after May 2011: see Figure 1). This phenomenon could allow social media to better approximate the opinion of the general public. While quite speculative, this conclusion, on its own, should be good news for the electoral forecasting ability of social media analysis. The following two sections explore more precisely this latter possibility.

Electoral campaign and social media (1): The 2012 French presidential ballot

We take an even more dynamic approach in our second example than in the previous case. By focusing on the second round of the 2012 French presidential elections held on 6 May 2012, when Sarkozy and Hollande underwent their final struggle, we tested whether the analysis of social media can be used as a device to forecast the actual results of the elections, comparing our results with those provided by traditional survey polls. We show the (unique) ability of social media analysis to monitor day-by-day the flow of Internet users' preferences as expressed by their tweets and their (close) connection with the ongoing political agenda and electoral campaign.

For this purpose, we collected 244,000 tweets posted between 27 April and 5 May. In the polls, Hollande won the ballot against Sarkozy with 51.64% of the total votes. According to the opinions expressed online, the night ahead of the election, we similarly foresaw a victory for the socialist candidate, Hollande, with 54.9% of votes. Moreover, our prediction was in line with predictions made by survey companies, who assigned a share of votes ranging between 52.5% (Ipsos) and 53.5% (TNS-Sofres) to Hollande in the last surveys published. Our estimate was also analogous to the prediction (53.2%) made by academic scholars (Nedeau et al., 2012).

As already mentioned, instead of running a unique analysis, during the run-off we continuously monitored the flow of preferences, day-by-day. We ran eight daily analyses to check how the expression of preferences changed over time in response to news related to the electoral campaign. Figure 2 displays the daily monitoring of voting preferences. Hollande almost always led by a narrow margin. However, this trend presented some peaks and turning points that could be explained in the light of the electoral campaign agenda.

First, on 28 April 2012, we detected a peak in favor of Hollande. At the time, Sarkozy was dealing with a document that seemed to attest that his electoral campaign in 2007 had been funded by the former ruler of Libya, Muammar Gaddafi. At the same time, another scandal involved the incumbent candidate: news reported by the media claimed that the popular socialist politician Dominique Strauss-Khan (DSK), a strong opponent of Sarkozy, had been illegally spied on by the French Secret Service. These stories created an echo online and contributed to the growth of support for Sarkozy's opponent, Hollande.

Conversely, in the following days (29 and 30 April 2012), a former member of the Libyan regime denied any suspicion about illegal funding in favor of Sarkozy, who in

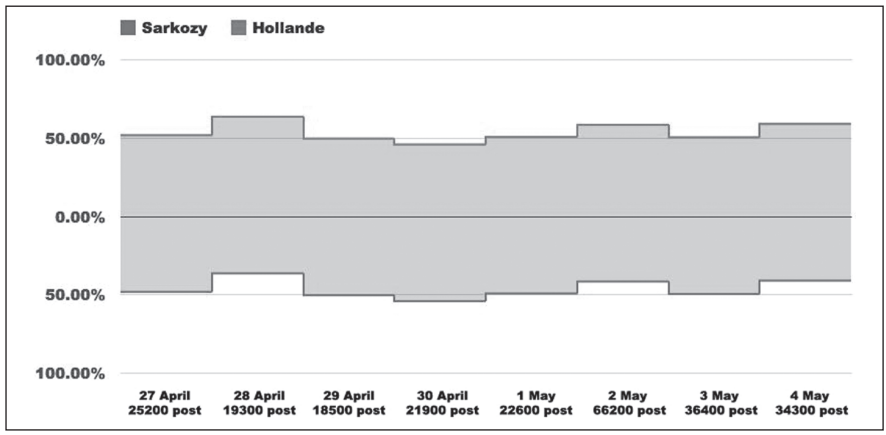

Figure 2. Flow of preferences expressed on Twitter during the electoral campaign for the second round of the 2012 French presidential election.

turn blamed the media for reporting fake news. In addition, DSK, who was involved in a sex scandal several months before the elections, entered the campaign and participated in a fund-raising dinner organized by members of the Socialist Party. As a consequence of these events, Sarkozy was able to gain support among the voters. However, he was not able to maintain a consensus. In fact, his provocative idea to celebrate the "real workers" day on 1 May was not fully supported by public opinion, leading to a loss of votes. Hollande's advantage grew even more after the televised debate held on 2 May 2012. This day was said to be a crucial event in the campaign, and in fact, we registered a large number of tweets written during or immediately after the debate (among the 66,200 tweets written on 2 May 2012, two thirds related to the debate). In line with other analyses, our estimates confirm that Hollande prevailed during the debate.[19] Just before the elections, the socialist candidate was safely leading. Finally, on the last day of the campaign, the centrist leader Bayrou granted his support to Hollande. His choice, however, seems to have pushed moderate voters to vote for Sarkozy, thus reducing the gap between the two candidates.

According to this analysis, we illustrated how voting preferences expressed day-by-day on Twitter tend to react to exogenous events that are related to the agenda of the electoral campaign. Furthermore, the number of preferences expressed in the week before the election enabled us to correctly forecast the outcome of the polls, yielding predictions that were very close to those made by traditional surveys.

Electoral campaign and social media (2): The 2012 French legislative elections

We also evaluated the predictive ability of social media analysis by applying this technique to forecast the outcome of the first round of the 2012 French legislative election, held on 10 June 2012. Compared to the previous case, this clearly represents a more

difficult (and more ambitious) challenge, given the large number of parties competing in that election. We gathered 79,300 tweets released during the week before the elections to predict the national share of votes of the main parties. As shown in Figure 3 below, at the national level, our prediction was once again close to the actual results. This outcome was true for almost every party. In particular, we made a very accurate forecast concerning the UMP, the Greens, the minor moderate parties and, to a lesser extent, the Socialist Party. On the contrary, we overestimated the support for far-left parties (FdG, NPA and others), and the National Front's (NF) share of the vote was underestimated. A possible explanation for our mis-estimation of the NF vote share is that far-right voters tend to be under-represented online (this phenomenon is particularly true for voters). In addition, it may be the case that NF voters are (more) reluctant to publicly express their voting behavior online. Similarly, left-wing voters seem to be over-represented on social networks, and this aspect could have led to inaccurate prediction.[20]

Nonetheless, on average, the Mean Absolute Error (MAE) of our prediction remains quite low, equal to 2.38%, which is not far from the MAE values displayed by the surveys held in the last week before the elections. On average, survey polls registered an MAE equal to 1.23%, ranging from 0.69% to 1.93%.

The data on the French legislative election allow for exploration of the possibility of assessing the main sources of bias that may alter the accuracy of our prediction. To do that, we use data on the local constituencies. We exploited the geo-tagging service made available through Twitter to gather preferences within 13 local areas:

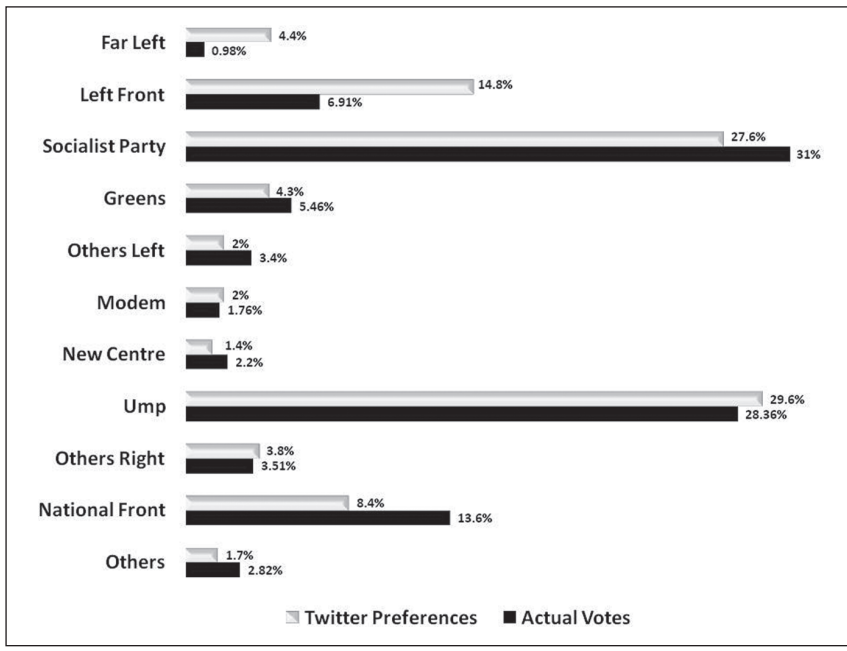

Figure 3. Predicted and actual vote shares related to the first round of the 2012 French legislative elections.

Bordeaux, Djion, Le Havre, Lille, Lyon, Marseille, Montpellier, Nice, Rennes, Saint Etienne, Strasbourg, Toulouse, and Toulon. We then ran 13 analyses to obtain the social media prediction within each area and compared these estimates with the actual results in the 46 local districts connected to those cities.[21] We measured the MAE, which represents our dependent variable and varies between 2.70 and 8.23, then tried to assess which elements increase or decrease the MAE of our prediction. We estimated three different models. Model 1 includes our main independent variables, *Number of Tweets*, the number of comments released in each area, which expresses the information available, and *Abstention*, the percentage of district voters who decided to abstain. In Model 2, we added three control variables: *Le Pen Votes Share*, the share of votes gained in the district by the far-right candidate during the 2012 presidential elections (used to identify those areas where the extreme right was the strongest); *Mélenchon Votes Share*, the share of votes gained in the district by the candidate of the Front de Gauche during the 2012 Presidential elections (as a proxy for the "red" districts), and *Incumbent*, a dummy variable equal to one when the incumbent MP was running to seek re-election. Finally, in Model 3, we added an interaction term between *Number of Tweets* and *Abstention* to assess whether the effect of having additional information about citizens' preferences is conditional on the likelihood that citizens will actually cast their vote. The data were analyzed using OLS regression. Table 2 reports the results.

From Model 1, we observed that any growth of the information available online improved our predictive skills. For instance, an increase of 1000 in the number of tweets analyzed lowered our error by approximately a quarter point. Conversely, the MAE was

Table 2. OLS regression of Mean Absolute Error.

Variables	*(1)*	*(2)*	*(3)*
Number of tweets	–0.000245**	–0.000234**	–0.004339***
	(0.000102)	(0.000108)	(0.001354)
Abstention	0.121490**	0.116903*	–0.227582*
	(0.054625)	(0.058571)	(0.125282)
Number of tweets × Abstention	–	–	0.000091***
			(0.000030)
Le Pen votes share	–	0.012887	0.001895
		(0.038121)	(0.034903)
Mélenchon votes share	–	–0.027611	–0.039590
		(0.104928)	(0.095634)
Incumbent	–	–0.383584	–0.635201
		(0.460141)	(0.427129)
Constant	1.028943	1.618636	17.68222
	(2.506913)	(2.828653)	(5.880315)
Observations	46	46	46
R-squared	0.191	0.210	0.361

Robust standard errors in parentheses.
***$p<0.01$,** $p<0.05$,* $p<0.1$.

greater when *Abstention* increased (the same concerns affect traditional offline pre-electoral polls: see Crespi, 1988). This scenario could occur because some citizens easily express their opinion online, while refusing to cast a vote (perhaps because they feel that their choice will not alter the results or because the act of voting is costly: Downs, 1957). Social media analysis seems less able to provide accurate predictions when voters tend to abstain at a higher rate (a 10% increase in *Abstention* increased the MAE by 1.2 additional points), while the accuracy should be greater when forecasting elections with a higher turnout. These two effects hold when adding certain control variables (Model 2). Our prediction does not therefore appear to be biased by the stronger presence of the NF or left-wing voters. While we know that incumbent candidates usually benefit from an advantage when seeking re-election (Ansolabehere and Snyder, 2002), it has also been argued that elections are a referenda on the incumbent (Freeman and Bleifuss, 2006), and these candidates may outperform in the pre-electoral surveys compared to the actual results due to name recognition. However, Table 2 shows that this potential "incumbency effect" does not damage our predictive ability.

Finally, in Model 3, we tested the conditional effect of *Number of Tweets* and *Abstention*. The coefficient of the interaction term was significant. Accordingly, in Figure 4, we report the marginal effect of *Number of Tweets* as the level of *Abstention* increases. We also superimpose a histogram portraying the frequency distribution for *Abstention* (the scale for the distribution is given by the vertical axis on the left-hand side of the graph). Having more information on citizens' voting preferences decreases the error only when the turnout rate is sufficiently high. Up to such threshold, our predictive skills are enhanced by any increase in the number of comments about voting

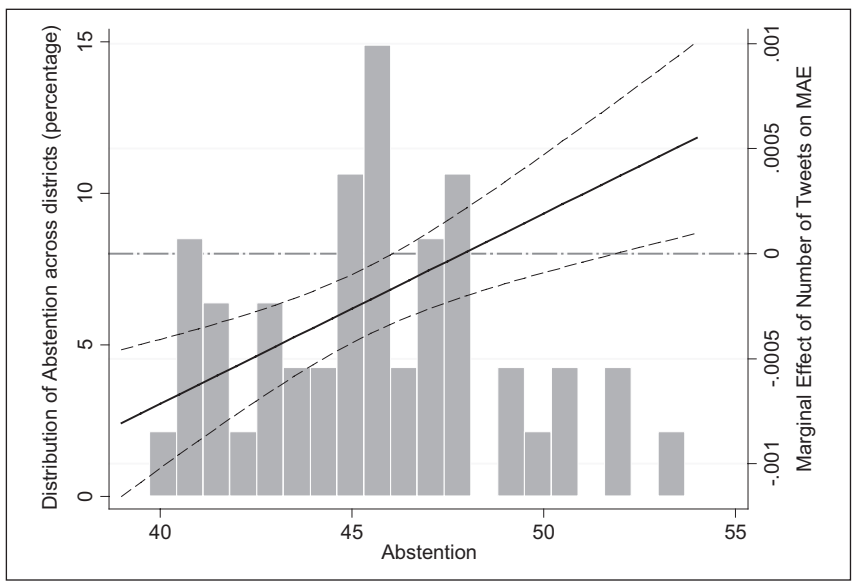

Figure 4. Marginal effect of the number of tweets on the Mean Absolute Error (with 90% confidence interval).

intentions, and such an effect increases as turnout grows. Conversely, when voters tend to abstain at a higher rate, having more information about their (declared) voting choice negatively affects the accuracy of our predictions: given that voters are more likely to express themselves on Twitter than to cast a real vote, the Mean Absolute Error tends to increase. This (original) finding is quite interesting, given that it clearly reflects the strict relationship between what happens online and offline in terms of our ability to extract reliable measures through social media analysis.

Conclusions

The growing usage of social media by Internet users who use social networking sites to express their opinions on a wide variety of topics, including their political and policy preferences, has raised interest in the possibility of exploiting this information to better understand the link between political preferences and political behavior. Not surprisingly, in recent years, there has been a dramatic increase in the number of works analyzing social media to assess the opinions of Internet users and to determine whether the attitudes expressed online can be used to forecast the voting behavior of the entire voting population. For all of these reasons, the ability to rely on techniques for measuring online public opinion has become an important topic.

In this paper, we have applied, in three different political scenarios, a statistical method that was recently introduced in the literature in order to perform supervised SA on blogs and social networks. This method improves on traditional SA, producing more accurate results. From the results of our empirical analyses, we can make some general claims. First, although Internet users are not necessarily representative of the entire population of a country's citizens, our analysis shows – with only few exceptions – a consistent correlation between social media results and the results we could obtain from more traditional mass surveys, as well as a remarkable ability for social media to forecast electoral results on average (a careful prediction that could not be simply due to chance).

This phenomenon seems to be true for both "single-issue" elections (such as a presidential race), in which the preference eventually expressed by an Internet user involves only a positive or a negative evaluation among two single options, and for situations which are more difficult to forecast, such as those in which Internet users can choose to express a preference among a (large) number of different *targets* (such as political leaders or political parties). This, together with the fact that the political scenarios analyzed here represented two different countries (Italy and France) in which the socio-economic-political traits of Internet users are not supposed to be necessarily identical, is clearly important for the robustness of our results.

The question of the direction of causality of this pattern (i.e., is the social media opinion becoming more similar to the general public opinion, or, on the contrary, are social media driving (or anticipating) the general public opinion?) lies beyond the scope of our research. However, this is clearly a topic that deserves further investigation.

In addition, the previous analyses also allow us to provide several more fine-grained conclusions. For example, SA of social media seems to provide more accurate predictions when focusing on the most popular leaders or on *mainstream* parties. However, the accuracy of predictions based on SA for *non*-mainstream parties could be increased by

developing an appropriate set of weights according to the political preferences of social media users (discounting, for example, the fact that in France, supporters of far-right parties tend to be under-represented on social networks compared to radical left-wing voters), provided this type of information is available (and reliable). This phenomenon represents another topic that needs deeper exploration.

Second, online preferences tend to react to exogenous factors (i.e., news, political agenda, electoral campaign) as expected (Franch, 2012), and these reactions seem to be in line with those observed through mass surveys.

Finally, some of the potential bias that arises from social media analysis may be softened in the long run as social network use increases: as we have shown, when a growing number of citizens express their opinions and/or voting choices online, the accuracy of social media analysis increases, provided that Internet users act in a way that is consistent with their statements online (for example, confirming their (declared) online preference by casting a (real) vote).

In summary, despite the well-known limits and troubles faced by social media analysis (Gayo-Avello et al., 2011), our results provide grounds to be optimistic about the capability of SA to become (if, indeed, it is not already) a useful supplement to traditional offline polls.

Funding

This research was supported by in-kind access to Crimson Hexagon's ForSight™ social media analysis platform through the Crimson Hexagon Social Research Grant Program.

Notes

1. Sentiment analysis consists of analyzing texts to extract information.
2. International Telecommunications Unions (http://www.itu.int/ITUD/ict/facts/2011/material/ICTFactsFigures2011.pdf).
3. Social Network around the World 2010 (http://www.slideshare.net/stevenvanbelleghem/social-networks-around-the-world-2010).
4. http://www.internetworldstats.com/facebook.htm.
5. Wasserman, Todd. March 21, 2012. "Twitter Says It Has 140 Million Users" http://mashable.com/2012/03/21/twitter-has-140-million-users.
6. For a more skeptical view of the role that social media can play in organizing revolts, see Morozov (2009) with respect to protests related to the 2009 Iranian elections.
7. During the EU elections held in 2009, the Pirate Party won 7.1% of votes in Sweden, gaining one seat in the EU parliament. In Germany, it received 2% of votes in the 2009 German Federal Election. It subsequently obtained positive results in German regional elections. In Italy, the *Movimento 5 Stelle* also reported surprising results during local elections held between 2009 and 2012, before winning a striking 25% of the votes in Italy's 2013 General Election.
8. A second (large) stream of research in the literature on social media adopts a more "political supply-side" approach, analyzing how the Internet and the diffusion of social media affect the content of electoral campaigns and political communication by candidates/parties (Larsson and Moe, 2012; Smith, 2009).
9. See http://www.crimsonhexagon.com/quantitative-analysis/. From our replications, the root mean square error of the estimates drops to 1.5% when the number of hand-coded documents is increased to 500.

10. In other words, a "word profile" is a vector composed of 0s and 1s: 0 is assigned when a term does not appear in the unit (but is used in some other units), and 1 is assigned when a term appears in the unit.
11. The population of tweets collected consists of all the tweets posted during the temporal period considered (see below) which include in their text at least one of a set of keywords (generally the name of the political leaders/parties covered by each of our analyses in both Italy and France).
12. Script and data are available upon request.
13. Note, however, that the Internet ceases to be a free environment for political debate whenever users are confronted with censorship, such as occurs in authoritarian regimes (King et al., 2012).
14. The "spiral of silence" theory claims that individuals who perceive their opinion to be in the minority do not tend to express their opinion, thereby strengthening the relative support for the (perceived) dominant views.
15. It has also been noted that the Internet tends to be dominated by a small number of heavy users who write more, while many users comment very rarely (Mustafaraj et al., 2011; Tumasjan et al., 2010). In addition, some accounts are false (Metaxas and Mustafaraj, 2010). Finally, and by definition, we can only observe the online opinions of those who have decided to express their attitudes (Gayo-Avello et al., 2011).
16. Gender constitutes a partial exception given that, despite equal participation in social networks, males tend to express their political views more than females do.
17. The survey question on which the popularity rating is based is the following: "I am going to read now a list of some political leaders. For each of them, please tell me if you have ever heard about him/her. If so, please tell me how would you judge him/her giving a score from 1 to 10: 1 meaning completely negative judgment, 10 completely positive judgment and 6 sufficiently positive."
18. The Berlusconi IV cabinet was weakened by the split of the *formateur* party, the People of Freedom (PDL), in 2010 (Ceron, 2011). Such weakness, exacerbated by the economic crisis and the striking growth of public debt, jeopardized government stability, paving the way for the anticipated elections. Although members of the ruling coalition argued for the dissolution of parliament after Berlusconi's resignation due to an escalation of the financial crisis, a majority of MPs surprisingly agreed to support a caretaker government led by the former EU commissioner, Mario Monti.
19. See http://tempsreel.nouvelobs.com/election-presidentielle-2012/20120502.OBS9654/sarkozy-ou-hollande-qui-gagne-le-debat-l-analyse-sur-twitter.html.
20. Alternatively, it could be the case that far-left parties tend to be more heavily affected by strategic voting in the first round, such that a (radical) left-wing Internet user expresses her sincere preference online but not at the polls. Although in a run-off electoral system, such as the one applied in the French legislative election, incentives to vote strategically are stronger in the second round, they are not absent in the first round (Cox, 1997). Note that such an incentive to express a sincere vote online and then to vote differently does not exist, by definition, when we have just two parties/candidates running at the polls. This could also explain why our estimations for the second round of the French presidential election appear slightly better than those for the French legislative election.
21. We excluded Paris due to its large size, which makes it more difficult to establish a link between the origin of each post and the electoral districts existing in the city.

References

Albrecht S, Lübcke M and Hartig-Perschke R (2007) Weblog campaigning in the German Bundestag election 2005. *Social Science Computer Review* 25(4): 504–520.

Alvarez RM and Hall TE (2011) *Electronic Elections: The Perils and Promises of Digital Democracy*. Princeton, NJ: Princeton University Press.

Ansolabehere S and Snyder JM (2002) The incumbency advantage in U.S. elections: an analysis of state and federal offices, 1942-2000. *Election Law Journal* 1(3): 315–338.

Bakker TP and De Vreese CH (2011) Good news for the future? Young people, Internet use, and political participation. *Communication Research* 20(10): 1–20.

Benkler Y (2006) *The Wealth of Networks*. New Haven, CT: Yale University Press.

Bennett WL and Segerberg A (2011) Digital media and the personalization of collective action. *Information, Communication & Society* 14(6): 770–799.

Best SJ and Krueger BS (2005) Analyzing the representativeness of Internet political participation. *Political Behavior* 27(2): 183–216.

Ceron A (2011) From words to facts: Wordfish, a modern technique to estimate policy positions of political actors. *Italian Political Science* 6. Available at: ow.ly/i8Mj4.

Chung J and Mustafaraj E (2011) Can collective sentiment expressed on Twitter predict political elections? In: *Proceedings of the twenty-fifth AAAI conference on artificial intelligence*, San Francisco, CA 7–11 August.

Cottle S (2011) Media and the Arab uprisings of 2011. *Journalism* 12(5): 647–659.

Cox G (1997) *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. New York: Cambridge University Press.

Crespi I (1988) *Pre-Election Polling: Sources of Accuracy and Error*. New York: Russell Sage.

De Zúñiga GH, Puig-I-Abril E and Rojas H (2009) Weblogs, traditional sources online and political participation. *New Media & Society* 11(4): 553–574.

Downey J and Fenton N (2003) New media, counter publicity and the public sphere. *New Media & Society* 5(2): 185–202.

Downs A (1957) *An Economic Theory of Democracy*. New York: Harper & Row.

Fox S (2010) Four in ten seniors go online. Available at: www.pewinternet.org/Commentary/2010/January/38-of-adults-age-65-go-online.aspx

Franch F (2012) (Wisdom of the Crowds)²: 2010 UK election prediction with social media. *Journal of Information Technology & Politics* 6(2): 87110.

Freeman SF and Bleifuss J (2006) *Was the 2004 Presidential Election Stolen?* New York: Seven Stories.

Gayo-Avello D (2011) Don't Turn Social Media Into Another 'Literary Digest' Poll. *Communications of the ACM* 54(10): 121–128.

Gayo-Avello D, Metaxas P and Mustafaraj E (2011) Limits of electoral predictions using social media data. In: *Proceedings of the international AAAI conference on weblogs and social media*, Barcelona, Spain 17–21 July.

Ghannam J (2011) *Social Media in the Arab World: Leading up to the Uprisings of 2011*. Washington, DC: Center for International Media Assistance.

Gloor PA, Krauss J, Nann S, et al. (2009) Web Science 2.0: identifying trends through semantic social network analysis. *International Conference on Computational Science and Engineering* 4: 215–222.

Goldstein P and Rainey J (2010) The 2010 elections: Twitter isn't a very reliable prediction tool. Available at: lat.ms/fSXqZW

Hindman M (2009) *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.

Hopkins DJ and King G (2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1): 229–247.

Jansen BJ, Zhang MM, Sobel K, et al. (2009) Twitter power: tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology* 60(11): 2169–2188.

Jansen HJ and Koop R (2005) Pundits, ideologues, and ranters: the British Columbia election online. *Canadian Journal of Communication* 30(4): 613–632.

Jungherr A, Jürgens P and Schoen H (2012) Why the pirate party won the German election of 2009 or the trouble with predictions. *Social Science Computer Review* 30(2): 229–234.

King G, Pan J and Roberts M (2012) How censorship in China allows government criticism but silences collective expression. Available at: gking.harvard.edu/files/censored.pdf

Langman L (2005) From virtual public spheres to global justice: a critical theory of Internetworked social movements. *Sociological Theory* 23(1): 42–74.

Larsson AO and Moe H (2012) Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society* 14(5): 729–747.

Lindsay R (2008) Predicting polls with Lexicon. Available at: languagewrong.tumblr.com/post/55722687/predicting-polls-with-lexicon

Madge C, Meek J, Wellens J, et al. (2009) Facebook, social integration and informal learning at university. *Learning, Media and Technology* 34(2): 141–155.

Mazzoleni G, Vigevani G and Splendore S (2011) *Mapping Digital Media: Italy*. New York: Open Society Foundations.

Metaxas PT and Mustafaraj E (2010) From obscurity to prominence in minutes: political speech and real-time search. Available at: bit.ly/h3Mfld

Morozov E (2009) Iran: downside to the 'Twitter revolution'. *Dissent* 56(4): 10–14.

Mustafaraj E, Finn S, Whitlock C, et al. (2011) Vocal minority versus silent majority: discovering the opinions of the long tail. In: *In: Proceedings of SocialCom/PASSAT,* Boston, MA, USA, 9–11 October 2011, pp. 103–110.

Nedeau R, Lewis-Beck MS and Bélanger E (2012) Proxy models for election forecasting: the 2012 French test. *French Politics* 10(1): 1–10.

Noelle-Neumann E (1974) The spiral of silence: a theory of public opinion. *Journal of Communication* 24(2): 43–51.

O'Connor B, Balasubramanyan R, Routledge BR, et al. (2010) From tweets to polls: linking text sentiment to public opinion time series. In: *Proceedings of the fourth international AAAI conference on weblogs and social media*, Washington, DC, 23–26 May.

Papacharissi Z (2002) The virtual sphere: the Internet as a public sphere. *New Media & Society* 4(1): 9–27.

Payne S (1951) *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.

Sang TKE and Bos J (2012) Predicting the 2011 Dutch Senate election results with Twitter. In: *Proceedings of SASN 2012, the EACL workshop on semantic analysis in social networks*, Avignon, France, 23–27 April.

Savigny H (2002) Public opinion, political communication and the Internet. *Politics* 22(1): 1–8.

Segerberg A and Bennett WL (2011) Social media and the organization of collective action: using Twitter to explore the ecologies of two climate change protests. *The Communication Review* 14(3): 197–215.

Smith A (2009) *The Internet's Role in Campaign 2008*. Washington, DC: Pew Research Center.

Tjong KSE and Bos J (2012) Predicting the 2011 Dutch Senate Election Results with Twitter. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 23–27 April 2012, pp. 53–60.

Tumasjan A, Sprenger TO, Philipp GS, et al. (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: *Proceedings of the fourth international AAAI conference on weblogs and social media* , Washington, DC, 23–26 May.

Upton G (2010) Does attractiveness of candidates affect election outcomes? Available at: com/lib/files/AttractivePoliticians.pdf

Véronis J (2007) Citations dans la presse et résultats du premier tour de la présidentielle 2007. Available at: aixtal.blogspot.com/2007/04/2007-la-presse-fait-mieux-que-les.html

Wei L and Hindman DB (2011) Does the digital divide matter more? Comparing the effects of new media and old media use on the education-based knowledge gap. *Mass Communication and Society* 14(2): 216–235.

Williams C and Gulati G (2008) What is a social network worth? Facebook and vote share in the 2008 presidential primaries. In: *Annual Meeting of the American Political Science Association*, Boston, MA, 28–31 August, pp. 1–17.

Woodly D (2007) New competencies in democratic communication? Blogs, agenda setting and political participation. *Public Choice* 134(1–2): 109–123.

Xin J, Gallagher A, Cao L, et al. (2010) The wisdom of social multimedia. In: *Proceedings of ACM multimedia 2010 international conference*, Firenze, 25–29 October, pp. 1235–1244.

Author biographies

Andrea Ceron, PhD in Political Studies, Università degli Studi di Milano. Research Fellow at Università degli Studi di Milano, Milan. His research focuses on intra-party politics, party competition, legislative studies, and social media analysis.

Luigi Curini, PhD in Institutions and Organizations, Università Cattolica di Milano. Associate professor of political science at Università degli studi di Milano (Italy). His research focuses on party competition, spatial theory of voting, and social media analysis.

Stefano M Iacus, PhD in Statistics at Università degli Studi di Padova (Italy), PhD in Mathematics at University of Le Mans (France). Associate professor of mathematical statistics and probability at the Università degli Studi di Milano. R Core developer team. Fields of interest include computational statistics, theoretical statistics, inference for stochastic processes, text mining and sentiment analysis, and causal inference.

Giuseppe Porro, PhD in Economics, Università degli studi di Pavia (Italy). Associate professor of economic policy at Università degli Studi dell'Insubria. He has taught at Università Bocconi, Milan; Università degli Studi di Trieste; and Università degli Studi di Milano. His main research fields are labor economics, policy evaluation, and economic dynamics.