Tweet your vote: How content analysis of social networks can improve our knowledge of citizens’ policy preferences. An application to Italy and France

Andrea Ceron, Luigi Curini, Stefano M. Iacus - University of Milan
Giuseppe Porro - Insubria University

To be presented at the XXVI Congress of Italian Society of Political Science (SISP), Rome, 13-15 September

Do not quote without authors’ permission

September 3, 2012

Abstract

The growing usage of internet and social media by a wider audience of citizens sharply increases the possibility to investigate the web as a device to explore and track their (policy) preferences. In the present paper we apply the recent method proposed by Hopkins and King (2010) to three very different scenarios, by tracking on one side the on-line popularity of Italian political leaders throughout 2011, and on the other side the voting intention of French internet-users in both the 2012 Presidential ballot and in the subsequent Legislative election. The variety of contexts so analyzed has been deliberately pursued to better investigate the strength and the limits of monitoring social networks, as well as to assess which factors can increase (or decrease) their reliability. Despite internet users are not necessarily representative of the whole population of country’s citizens, our analysis shows a remarkable ability of social-media to forecast electoral results as well as a consistent correlation between social-media results and the ones we obtain from more traditional mass surveys. Still, sentiment analysis of social network seems to provide more accurate predictions when focusing on the most popular leaders or on ‘mainstream’ parties, while both traditional surveys as well as social networks analyses methods appear to be affected, at least in part, by the same sources of bias, like the strategic behavior and the spiral of silence. Finally, the possibility of an ‘information overload’ arises as a further factor affecting negatively the analysis of social-media.
Introduction

The exponential growth of social media and social networks, like Facebook and Twitter, raises the possibility to delve into the web to explore and track the (policy) preferences of citizens and voters. Internet in fact represents a valuable source of data useful to monitor public opinion (Madge et al. 2009; Woodly 2007) and thanks to the recent developments of quantitative text analysis and blog sentiment analysis (BSA) we are now in a better position to exploit such information in a reliable manner.

As a matter of fact, scholars have recently started to explore social media as a device to forecast the elections (Tjong and Bos 2012), to assess the popularity of politicians (Gloor et al. 2009), or to compare the citizens’ political preferences expressed on-line with those caught by traditional polls (O’Connor et al. 2010). Some of these works rely on very simple techniques, focusing on the volume of data related to parties or candidates. For instance, Véronis (2007) proved that the simple number of candidate mentions in blog posts is a good predictor of electoral success and can perform better than election polls. Along the same line, some scholars claimed that the 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 relative number of tweets related to each party is once again a good predictor for its vote share.

Still not all enquiries succeeded in correctly predicting the outcome of the elections (Gayo-Avello et al. 2011; Goldstein and Rainey 2010). For instance, it has been shown that the share of campaign weblogs prior to the 2005 federal election in Germany was not a good predictor of the relative strength of the parties insofar as small parties were overrepresented (Albrecht et al. 2007). In a study about Canadian elections, Jansen and Koop (2005) failed in estimating the positions of the two largest parties that were reversed. Gayo-Avello (2011) proved that social media analysts would have overestimated Obama’s victory in 2008 (up to the point of predicting his success even in Texas). Finally, Jungherr et al. (2011) criticized the work of Tumasjan et al. (2010) arguing that it does not satisfy the alternative from irrelevant alternatives criterion (e.g., including the Pirate Party into the analysis would have had yielded a negative effect on accuracy of the predictions).

It has been also noted that the mere count of mentions or tweets is not sufficient to provide an accurate and reliable foresight (Chung and Mustafaraj 2011). Accordingly, other studies tried to improve this stream of research by means of sentiment analysis. Lindsay (2008), for example, built a sentiment classifier based on lexical induction and found correlations between several polls conducted during the 2008 presidential election and the content of wall posts available on Facebook. O’Connor et al. (2010) have shown similar results displaying correlation between Obama’s approval rate and the sentiment expressed by Twitter users. In addition, sentiment analysis of tweets proved to perform

---

1Sentiment analysis consists in analyzing texts to extract information. Basic sentiment analysis allows to determine the attitude of the author measuring the polarity (negative or positive) of such sentiment with respect to one particular topic.
as well as polls in predicting the results of 2011 Dutch Senate Election (Tjong Kim Sang and Bos 2012) and the analysis of multiple social media (Facebook, Twitter, Google and YouTube) was able to outperform traditional surveys in estimating the results of the 2010 UK Election (Franch 2012).

In the present paper we follow this latter route by adopting the method proposed by Hopkins and King (2010) (HK since now). As we will discuss, this method presents various advantages compared to traditional BSA techniques. We will employ it in three very different scenarios, by tracking on one side the on-line popularity of Italian political leaders throughout 2011, and on the other side the voting intention of French internet-users in both the 2012 Presidential ballot and in the subsequent Legislative election. In all cases we will contrast our results with the ones obtained through traditional off-line surveys as well as to the actual electoral results.

The variety of contexts so analyzed has been deliberately pursued to better investigate the strength and the limits of monitoring social networks, as well as to assess which factors can increase (or decrease) their reliability. In the conclusion we advance some suggestions for future research.

1 Social-Media: a seemingly (striking) growing role

Nowadays, Internet access is available to a wider audience of citizens (and voters). In turn, the usage of social media is growing at very fast rates. Around 35 people out of 100 got access to the web, all over the world, in 2011. This means that the share of internet users is greater than one third of the world population (approximately 2.5 billion people). Many of them pay attention to social medias and make use of social networks, like Facebook (over 800 million of users, 12% of the world population) or Twitter (140 million of active users). Overall, 72% of the internet population is active on at least one social network.

The huge number of people connected to social network is by itself a good reason to recall the importance of analyzing social media, albeit by far not the only one. Recently, in fact, social network have started to wield substantive effects on real world politics. We can remind, for instance, how social media have been used to organize revolts, particularly during the ‘Arab spring’, in Egypt, Libya, Syria, Tunisia (Ghannam 2011) or demonstra-

---

3Source: http://www.internetworldstats.com/facebook.htm
6Other scholars have however deemphasized the role played by Twitter and Facebook in organizing revolts showing that, for example, the percentage of social media users in Middle-Eastern countries is negligible (see Morozov 2009 with respect to protests related to the 2009 Iranian elections).
tions (like the one organized by the Italian feminist movement ‘Se non ora quando?’).  

In addition, social media and social network are nowadays widely used to build social movements and political parties, like the German Pirate Party or the Italian ‘Movimento 5 Stelle’, which uses the web to set the party line and to select candidates. Social media are also often used to present petitions and to complain against politicians’ decisions. In Italy, for example, the importance of the opinions expressed on-line has been recognized by the Monti cabinet, who decided using the web to ask the opinion of Italian citizens and to gather suggestions on two topics: the spending review and the legal value of the university degree.

In this sense, scholars claim that social media can improve political participation (both on-line and off-line). They show how internet usage and involvement in social media are positively related with various indicators of civic engagement (Anduiza et al. 2009; Boulliane 2009; Gil De Zuniga et al., 2009; Kaufhold et al., 2010; Östman 2012), albeit other studies show that the democratically desirable consequences of internet use apply mainly to those who are already engaged (e.g., Bimber 2001). Furthermore, on-line communities can also radicalize the positions of the users, rather than moderate them (Alvarez and Hall 2011; Hindman 2009).

A second (large) stream of research in the literature on social-media adopts a more political ‘supply-side’ approach, analyzing how internet and the diffusion of social-media has affected the content of the electoral campaigning and the political communication by the candidates and parties. While some of the initial hope for e-democracy have been unfulfilled (Chadwick 2008; Hilbert 2009), internet still provides new opportunities linked with electoral campaign (Larsson and Moe 2012; Smith 2009) such that politicians can engage with the wider public (Gibson et al. 2008). However the extent of these potential advantages should not be exaggerated given that politicians participation presents some limits (Vaccari 2008) and their on-line engagement follows a top-down pattern, restricting

---

7 ‘Se non ora quando?’ was the name of an Italian feminist movement raised in Spring 2011, after a sexual scandal involving the former Prime Minister, Berlusconi. This movement organized a non-partisan demonstration to protest against the government and to call for more gender equality in public life.

8 During the EU elections held in 2009, the Pirate Party won 7.1% of votes in Sweden, gaining 1 seats 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 winning 8.9% of votes in Berlin, 8.3% in Schleswig-Holstein, 7.5% in Nord Reno Vestfalia and 7.4% in Saarland. In Italy, the Movimento 5 Stelle reported surprising results during local elections held between 2009 and 2012. During the 2010 Regional Elections it received 6% of votes in Emilia-Romagna and 4% in Piemonte. After that it gathered 14% of votes in the elections for the mayor of Genova and 19.5% in the first round of the election in Parma, where its candidate has been elected as the mayor after the run-off. According to the survey polls, 20% of Italian voters are going to support such party for the 2013 Italian General Election.

9 http://www.corriere.it/economia/12_maggio_02/appello-governo-ai-cittadini-segnalazioni-sprechi-spending-review_9e5918c2-9438-11e1-ae3e-f83a8e51ff45.shtml

10 http://archiviostorico.corriere.it/2012/marzo/03/ Laurea_via_alla_consultazione_online_co_9_120303011.shtml

11 Interestingly, this last point about the impact of internet on the direction of citizens’ preferences, (i.e., making them more radical or more moderate) reminds pretty closely the heated debate in the deliberative democracy literature on the possible consequences of deliberation (Sunstein 2002).
the field for an open debate (Jackson and Lilleker 2009).

2 How to scrutinize citizens’ preferences through social-media: a novel method (and some pros and cons)

Given the wide amount of data about public opinion available on-line (and its growing relevance), monitoring this flow of preferences becomes an important task per-se. The problem is to select the kind of method more 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 catch the attitude of internet users going beyond the mere number of mentions. To this aim we will employ the method recently proposed by Hopkins and King (2010). The main advantage of the HK method is that it allows to perform an automated and at the same time supervised sentiment analysis. Such analysis is based on a two stage process. The first step involves human coders and consists in coding a subsample of the documents downloaded from some Internet source. This subsample represents a training set that will be used by the HK algorithm to classify all the unread documents, in the second stage. This method allows to overcome one of the major problems of the ‘old fashioned’ BSA based on ontological dictionaries, that is the fact that they tend to misclassify posts that do not adopt a straightforward language. On the contrary, human coders can be more precise during such stage (Hopkins and King 2012). Then, the automated statistical analysis provided by the HK algorithm extends such accuracy to the whole population of posts allowing to properly catch the opinions expressed on the web. For example, with only a hundred of hand-coded documents, the root mean square error of the estimates is around 3% (still lower than many survey polls), while it drops until 1.5% when we increase the number of hand-coded documents up to 500.

The methodology is based on the assumption that the opinion of people posting on social networks can be deduced by all the terms they use: not only the terms explicitly related to the topic they talk about, but also the ‘neutral’ part of the language commonly used. Therefore, in order to characterize the different opinions, the single units (blogs, tweets) in the data set are decomposed into single words or stems: consequently, each unit is represented by the vector of the terms used.

The formal background of the method is simple (see Hopkins-King 2010 for details). Indicate by $S$ the word profiles (i.e., all the stems appearing in a sentence) used in the text units and by $D$ the opinions expressed by people posting the texts. The frequency distribution of the terms $P(S)$ can be expressed as:

---

12For example, a sentence such as “this is a wonderful rubbish thing to say” could create a problem of classification for an ontological dictionary, given that it includes both a positive and a negative reference. However, for a human coder the meaning of the sentence (and its underlying attitude toward the subject of the sentence) is immediately evident.
where $P(D)$ is the frequency distribution of the opinions. The aim of the method is to get an estimate of $P(D)$, i.e. to know how the opinion is distributed over the posting population. The frequency distribution $P(S)$ can be evaluated tabulating all the texts posted, and it requires only some computer time and no debatable assumption. The conditional distribution $P(S|D)$ cannot be observed, and must be estimated by the hand-coding of the training set of texts.

The hand-coding of the training text, in fact, allows for calculating $P_T(S|D)$, i.e. the conditional frequency distribution of word profiles inside the training set. The assumption and the reasonable requirement - of the method is that the texts of the training set are homogeneous to the whole data set, i.e. they come from the same ‘world’ the rest of the dataset comes, such that one can assume that:

$$P_T(S|D) = P(S|D)$$ (2)

If this is the case, the frequency distribution of the opinions can be consistently estimated, because both $P(S)$ and $P_T(S|D)$ are observable:

$$P(D) = P(S)/P(S|D) = P(S)/P_T(S|D)$$ (3)

It is worth remarking that while the homogeneity of the training set to 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.13

Broadly speaking, there are several social-media that could be analyzed. Here, with a few exceptions, we will focus on Twitter, a social network for microblogging (Jansen et al. 2009) that experienced a sharply growth in the last months. Twitter is today the third-highest-ranking social network, behind Facebook and MySpace, whereas it ranked twenty-second only two years ago, in 2009. When it comes to the countries analyzed in this work we observe that one internet user out of two participates in social networks

---

13Besides, while classic web analyses allow only distinguishing between positive and negative references to a particular topic, the HK enables to measure also the intensity of preferences. In addition it also allows catching the motivation linked to people’s choice according to the variety of latent dimensions that emerge from the stream of voices available on the net.
(48.1% in Italy and 54% in France).\footnote{Source: eMarketer (http://www.emarketer.com/Article.aspx?R=1009019ecid=a6506033675d47f8881651943c21e5ed4)} In June 2011 Twitter was the second ranking in France and the third most-used social network in Italy. In particular, in February-March 2012, 12 millions of Italian users were active on Twitter, that means approximately 20% of the Italian population. Among them one out of three (6% of the whole population) released political comments and opinions.\footnote{Source: Osservatorio Politico ISPO (Istituto per gli Studi sulla Pubblica Opinione). We thank ISPO for having shared the data.} A further crucial advantage of Twitter, that makes it so popular in the literature on social-media analysis, is that all the posts by users (‘tweets’ in the Twitter jargon) can be freely accessible, contrary to other social-networks. Besides, an internal search engine in Twitter allows the researcher to select those tweets relevant for the topic under analysis.

To download all the data employed in the present paper we have relied on two Internet engines: Crimson Hexagon(http://www.crimsonhexagon.com/) and Voices from the Blogs (http://www.voicesfromtheblogs.com). In the Italian case the collection and the analysis of the tweets have been run using the ForSight platform provided by Crimson Hexagon,\footnote{We acknowledge the research grant obtained by Crimson Hexagon in this regard.} while in the French two cases the data have been collected through Voices from the Blogs and the analysis have been run in R.\footnote{Script and data available upon request} In both situations, however, the HK algorithm has been employed, so all analyses are perfectly comparable.

Compared to traditional survey polls, running an analysis on social-media is attractive for a number of reasons (Xin et al. 2010). First of all, social-media analysis is cheaper and faster compared to traditional surveys while enabling to continuously monitor public opinion performing a real time analysis (as well as to run, eventually, a retrospective analysis, by getting the opinion when it was actually expressed). On the contrary, off-line surveys are by definition more static. This feature ends up being very relevant during the electoral campaigns, as we will discuss below. In fact, thanks to the BSA, we can measure voters’ attitude day-by-day (at the extreme, also hour-by-hour). Hence we are able to catch the reaction of public opinion to any exogenous stimulus by observing the shift in preferences measured immediately after the shock. Along the same vein, analyzing social-media also allows to observe trends and breaking-points. This feature can have obvious implications for both researchers and spin doctors. Scholars can benefit from such amount of information to investigate preferences in-the-making, while analysts and advisors can exploit these data to adjust the frame of the electoral campaign.

In addition, traditional surveys pose solicited questions and it is well known this might inflate the share of strategic answers (Payne 1951). Conversely sentiment analysis does not make use of questionnaires and just focus on listening to the stream of unsolicited opinions freely expressed on internet. In other words it adopts a bottom-up approach, at least if compared with the more traditional top-down approach of off-line surveys. Far from saying that all the comments posted on social networks contain the
sincere preference of the author, it has been however often underlined that Internet represents to a large extent a public sphere where the debate is uncoerced by any established political agenda because users are free to express themselves (Savigny 2002).\textsuperscript{18} In this vein, social network should be in the position to be less affected by the spiral of silence (Noelle-Neumann 1974),\textsuperscript{19} or at least as biases as traditional polls. In fact, while web-analysis has to deal with the problem of silent users, surveys face the problem of low response rate.

The main weakness usually advanced when talking about social-media analysis is related to the riddle of being the ‘universe’ so analyzed representative or not of the whole population. Indeed, although the number of users is strikingly growing, the socioeconomic traits of citizens who have access to the web do not exactly match the actual demographics of the whole population of citizens and voters (Tjong Kim Sang and Bos 2012): for example, previous studies shown that senior citizens are underrepresented on the web (Fox 2010) and there is a prevalence of highly educated male individuals (Wei and Hindman 2011).\textsuperscript{20} Interestingly, the social-demographic differences are lowered when considering only the sample of people who release political opinions on-line (on this point, see also: Bakker and de Vreese 2011).\textsuperscript{21} Along this vein, Best and Krueger (2005, 204, italic added) shown 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”. On the other side, a review of queries related to the Italian case revealed that right-wing citizens are underrepresented, while the share of left-leaning people that are active on-line is greater. The same happens for certain geographical areas (the South) compared to others (the North-East). Finally, other studies claimed that different rates of internet availability do not wield strong effects on the electoral outcomes. Falck et al. (2012) found that an increase in the availability of DSL connection tend to decrease voter turnout, while it does not systematically benefit single parties. It only plays a negligible negative effect on the share of votes won by right-wing parties wielding no effect on the electoral fortunes of the left.

Summing up, although the social network population, so far, is not representative of one country’s citizenry, there are still some doubts about whether such bias could affect the predictive skills of social network analysis compared to traditional off-line surveys.

\textsuperscript{18}Note however that internet ceases to be a free environment for political debate whenever users have to deal with censorship, like in authoritarian regimes (King \textit{et al.} 2012).

\textsuperscript{19}The ‘spiral of silence’ theory claims that individuals who perceive their opinion to be in the minority do not tend to express such opinion at all, thereby strengthening the relative support for (perceived) dominant views.

\textsuperscript{20}It has also been noted that Internet tends to be dominated by a small number of heavy users that writes more, while many users make comments very seldom (Mustafaraj \textit{et al.} 2011; Tumasjan \textit{et al.} 2010). In addition, some accounts are fake (Metaxas and Mustafaraj 2010) and social media allow manipulation by spammers and propagandists. Finally, on-line we can observe only the opinion of those who have decided to express their attitudes (Gayo-Avello \textit{et al.} 2011).

\textsuperscript{21}Gender constitutes a partial exception given that, despite equal access to social networks, males tend to express their political views more than females do; source: ISTAT (http://www.istat.it/it/archivio/48388)
Indeed, the former aspect (the predictive skills of social-media analysis) does not necessarily need the previous factor (i.e., the issue of representation) to hold true to effectively applies.

This can happen, for example, if we assume that political active internet users act like opinion-makers that are able to influence (or to ‘anticipate’) the preferences of a wider audience: as a result, it could happen that the preferences expressed through social networks and social medias today will affect the opinion of the whole population tomorrow (O’Connor et al. 2010).

In the next sections we will therefore test the predictive skills of social network analysis by employing the HK method in two different countries (Italy and France) and over three distinguished political phenomena: leaders popularity, Presidential and Legislative national elections.

3 Comparing Italian leaders’ popularity ratings in 2011

Our first political context in which exploring the usefulness (and the reliability) of a social-media analysis concerns the relationship between the popularity ratings of the main Italian political leaders throughout 2011 as they arise from traditional Mass-Surveys (source: ISPO, Istituto per gli Studi sulla Opinione Pubblica) and from the analysis of social-media posts. We treat in this sense the former surveys as our benchmark, and we control how much the latter approach them.

The Mass-Surveys popularity ratings go from the 13th of January to the 20th of October 2011, and they focus on seven leaders: Silvio Berlusconi (leader of PDL and Italian Prime Minister at that time), Pier Luigi Bersani (leader of PD, main opposition party), Umberto Bossi (leader of Northern League and main cabinet partner of PDL at that time), Pier Ferdinando Casini (leader of the centrist party UDC, opposition party), Antonio Di Pietro (leader of IDV, opposition party), Gianfranco Fini (President of the Italian Lower Chamber and co-founder of PDL, before leaving the party at the end of 2011), and Nichi Vendola (leader of the radical-left party SEL, the main extra-parliamentary opposition party). The popularity ratings ranges from 0 to 100 and identify the percentage of positive scores given by the respondents to each leader.

Similarly, the popularity of each leader in the social media has been estimated as the percentage of his/her positive posts over the sum of his/her positive and negative posted, and once again ranges from 0 to 100 to make it comparable to the survey popularity ratings. We considered two different temporal range: in the first case, we collected all the posts concerning each leader in the month preceding the day in which the Mass-Survey was actually administered. In the second case, we rerun the above procedure considering just the week preceding the day in which the Mass-Survey was administered. Overall, we have analyzed over 107,000 tweets if we consider the monthly timing (32,000 tweets in the weekly timing). Given that the results of our analysis look remarkably similar regardless of the time-period considered when analyzing social-media, we focus here on
the popularity scores that arise from a weekly-timing (following the choice made by Tjong Kim Sang and Bos 2012).

In Table 1 below we report the average difference (Mass-Surveys minus Social-Media popularity ratings) of our scores so obtained. Three main findings clearly arise from the Table. First, as long as we consider all leaders without any internal distinction, the average Mass-Surveys ratings appear to be always higher than social-media ones (on average by 5 points). This is true for all the leaders, except Di Pietro and Fini whose on-line popularity appears to be higher. Second, we find a considerable variation among leaders: for example, the average difference between the two measure 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 Mass-Surveys and social-media ratings is positive, albeit not dramatically strong. Note, however, a marked contrast between, on one side Berlusconi, Bersani and Bossi (that is, the three most important and most visible Italian leaders during 2011) and on the other all the remaining leaders. For our first set of leaders the correlation is 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.

<table>
<thead>
<tr>
<th>Leader</th>
<th>% Positive-posts</th>
<th>Avg. difference</th>
<th>St.Dev.</th>
<th>r</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlusconi</td>
<td>8.71</td>
<td>1.89</td>
<td>.933</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Bersani</td>
<td>12.53</td>
<td>1.40</td>
<td>.746</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Bossi</td>
<td>2.58</td>
<td>2.54</td>
<td>.540</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Casini</td>
<td>13.34</td>
<td>2.92</td>
<td>-.008</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Di Pietro</td>
<td>-4.12</td>
<td>1.97</td>
<td>.109</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Fini</td>
<td>-2.30</td>
<td>2.93</td>
<td>.005</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Vendola</td>
<td>10.32</td>
<td>2.14</td>
<td>.090</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>All leaders</td>
<td>5.71</td>
<td>.497</td>
<td>.241</td>
<td>43</td>
<td></td>
</tr>
</tbody>
</table>

The previous Table gives us however just an aggregate (and static) picture that summarizes all the temporal observations. Therefore, it 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)\(^{22}\) of the predictions on leaders popularity as they arise from social-media as compared to the scores obtained from Mass-surveys. As can be seen, despite being quite relevant at the beginning of 2011 (around 13 points), this absolute difference tends to markedly decrease by more than the half as times goes by.

To sum up, albeit being just an exploratory analysis, our results provide some quite interesting insights. First, at least for the most visible leaders, the two measures of pop-

---

\(^{22}\)The MAE has been widely used to compare the accuracy of forecast based on social network analysis (Tumasjan *et al.* 2010) and that of political information markets relative to election polls (Huber and Hauser 2005).
Figure 1: The MAE of leaders’ popularity ratings between Mass-Surveys and Social-media estimation over time
ularity ratings (Mass-Surveys and social-media) seem to go hand-by-hand, 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, so it happens also with the second one and vice versa.

Second, 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 (absolute) average difference between the two measure of popularity ratings at least during 2011 seems to be clearly declining over time. It is worth noting that the possibility of new elections was a widely debate (and a real possibility) during the second part of 2011 in the Italian political debate throughout the political crisis of the Berlusconi IV cabinet. 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 since May 2011). This, on the other side, could turn social media to be able to better approximate the opinion of the general public. Albeit quite speculative, this conclusion, by its own, should be a good news for the electoral forecasting ability of a social-media analysis. The next two sections are devoted to explore precisely this latter possibility.

4 Electoral campaign and social media (1): the 2012 French presidential election ballot

In our second example, we take an even more dynamic approach 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 fought their final struggle, we test whether the analysis of social media can be a device to forecast the actual results of the elections, comparing our results with those provided by traditional survey polls. Secondly, we show the (unique) possibility that a social-media analysis allows to do, that is, the chance to monitor day-by-day the flow of preferences of internet users as expressed by their tweets and their (close) connection with the on-going political agenda and electoral campaign.

For this purpose we collected 244,000 tweets, posted between April 27th and May 5th. At the polls Hollande won the ballot against Sarkozy with 51.64% of total votes. According to the opinions expressed on-line we foresaw similarly a victory for the socialist candidate, Hollande, with the 54.9% of votes. Even considering the maximum error attached to our technique (discussed earlier) we are therefore able to correctly predict

The Berlusconi IV cabinet was weakened by the split of the formateur party, People of Freedom (PDL), in 2010 (Ceron 2011a, 2011b). After that, the ruling coalition composed by PDL and Northern League, was defeated during local elections and national referenda held in May and June 2011 (Chiaramonte and D’Alimonte 2012; Marangoni 2011). Such weakness, exacerbated by the economic crisis and the striking growth of public debt, jeopardized government stability paving the way for anticipated elections. Although members of the ruling coalition claimed for parliament dissolution after Berlusconi resignation, due to an escalation of the financial crisis, a majority of MPs surprisingly agreed on supporting a caretaker government led by the former EU commissioner, Mario Monti.
the outcome. Our prediction moreover is in line with those made by survey companies, who assigned to Hollande a share of votes ranging between 52% and 53.5% in the last published surveys. Our estimate was also analogous to the prediction (53.2%) made by academic scholars (Nedeau et al. 2012) and it was very close to the actual results too (51.6%).

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 analysis to check how the expression of preferences has changed over time, in reply to the news related to the electoral campaign. Figure 2 displays the daily monitor of voting preferences. Hollande was almost always leading though with a narrow margin. However the flow is not always straight and it highlights some peaks and turning points that could be explained in the light of electoral campaign agenda.

Figure 2: Flow of preferences expressed on Twitter during the electoral campaign for the second round of 2012 French Presidential elections

First of all, on 28th April we found a peak in favor of Hollande. In those days Sarkozy was dealing with a document who seemed to attest that his electoral campaign in 2007 was founded by the former ruler of Lybia Muammar Gaddafi. On the same days another scandal involved the incumbent candidate: a piece of news reported by the medias claimed that the popular socialist politician Dominique Strauss-Khan (DSK), a strong opponent of Sarkozy, has been illegally spied by the French Secret Service. These elements

For instance, Ipsos predicted his victory with 52.5% and TNS-Sofres with 53.5%.
had an echo on-line and contributed to a growth of support for Sarkozy’s opponent, Hollande.

Conversely, in the following days (the 29th and the 30th of April) a former member of the Libyan regime denied any suspicion about illegal funding in favor of Sarkozy, who in turn blamed the medias for reporting fake news. In addition DSK, who was involved in a sex scandal few months before the elections, entered into the campaign and took part to a fund-raising dinner organized by members of the Socialist Party. In consequence of these events Sarkozy was able to gain support among the voters. However he was not able to keep such consensus. In fact, his provocative idea to celebrate the ‘real workers’ day on the 1st of May did not reach the agreement of public opinion, wielding a loss of votes. Hollande advantage grew even more after the TV debate, held on May 2nd. This day was said to be a crucial event of the campaign, and in fact we registered a huge number of tweets written during or immediately after the debate (among the 66,200 tweets written on May 2nd, two third were comments to the debate). In line with other analyses our estimates confirm that Hollande has prevailed during the debate. Then, few days before the elections, the socialist candidate was safely leading. Finally, during the last day of campaign the centrist leader, Bayrou, granted his support to Hollande. His choice however seems to have pushed moderate voters to vote for Sarkozy reducing the gap between the two candidates.

According to this analysis we proved that voting preferences expressed day-by-day on Twitter tend to react to exogenous events related to the agenda of the electoral campaign. Furthermore, the amount of preferences expressed in the last week before the elections enables to correctly forecast the outcome of the polls, yielding predictions that are very close to those made by traditional surveys (see also: Larsson and Moe 2012; Tjong Kim Sang and Bos 2012).

5 Electoral campaign and social media (2): the 2012 French Legislative elections

We double checked the predictive skills of a social-media analysis by applying this technique to forecast the first round of the 2012 French Legislative election, held on 10 June 2012. Compared to the previous case, this represents clearly an harder (and more ambitious) exercise, given the large number of parties competing in that election. We gathered in this respect 79,300 tweets released during the last 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 is once again close to the actual results. This is true for almost every party. In particular we made a very accurate forecast concerning UMP, Greens, minor moderate parties and, to a lesser extent, the Socialist Party. On the contrary, we overestimated far left parties (FdG, NPA and others) while the vote share of National

---

Front has been underestimated. A possible explanation for misestimating the FN vote share is that far-right voters tend to be underrepresented on-line (this is particularly true for elder voters). In addition it can be that FN voters may be (more) reluctant to publicly express their voting behavior on-line. Similarly, left-wing voters seem to be overrepresented in social networks and this aspect could have led to inaccurate prediction.

Having noted that, on average the Mean Absolute Error (MAE) of our prediction remains quite low being equal to 2.38%, which is not far from those displayed by the surveys held during the last week before the elections. On average survey polls registered a MAE equal to 1.23%, ranging from 0.69% to 1.93%.

Figure 3: Predicted and actual votes share related to the first round of the 2012 French Legislative elections

---

Alternatively, it could be that the far-left parties tend to be the ones more heavily affected by strategic voting by voters in the first round, so that a (radical) left-wing internet user expresses her sincere preference on-line, but not at the polls. Albeit in a run-off electoral system, such as the one applied in the French Legislative election, the incentives to vote strategically are stronger in the second round, they are not absent also in the first round (Cox 1997). Note that such incentive to express a sincere vote on-line 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 the French legislative election case.

---

26 Alternatively, it could be that the far-left parties tend to be the ones more heavily affected by strategic voting by voters in the first round, so that a (radical) left-wing internet user expresses her sincere preference on-line, but not at the polls. Albeit in a run-off electoral system, such as the one applied in the French Legislative election, the incentives to vote strategically are stronger in the second round, they are not absent also in the first round (Cox 1997). Note that such incentive to express a sincere vote on-line 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 the French legislative election case.
The data on the French Legislative election allow us also to explore a further point, linked to the possibility to assess the main sources of bias that may alter the accuracy of our prediction. To do that we use data about local constituencies. We exploited the geo-tagging service made available through Twitter to gather preferences within 13 local areas: Bordeaux, Dijon, Le Havre, Lille, Lyon, Marseille, Montpellier, Nice, Rennes, Saint Etienne, Strasbourg, Toulouse, Toulon. Then we ran 13 analysis to get social-media prediction within each area and we compared such estimates with the actual results in the 46 local districts connected to those cities. We measured the MAE, which represents our dependent variable and varies between 2.70 and 8.23. Then we tried to assess which elements increase or decrease the mean absolute error of our prediction. Our independent variables are: Number of Tweets, the number of comments released in each area; Number of Tweets (Quadratic), which is the quadratic term; 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 is strongest); Abstention, the percentage of district voters who decided to abstain; District Population, the number of eligible voters within each district. Data have been analyzed through OLS. Table 2 reports the results.

Table 2: OLS regression of Mean Absolute Error

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Tweets</td>
<td>-0.00173***</td>
<td>(0.000520)</td>
</tr>
<tr>
<td>Number of Tweets (Quadratic)</td>
<td>1.85e-07***</td>
<td>(5.79e-08)</td>
</tr>
<tr>
<td>Le Pen Votes Share</td>
<td>0.0846**</td>
<td>(0.0346)</td>
</tr>
<tr>
<td>Abstention</td>
<td>0.0918*</td>
<td>(0.0533)</td>
</tr>
<tr>
<td>District Population</td>
<td>-2.28e-05***</td>
<td>(7.75e-06)</td>
</tr>
<tr>
<td>Constant</td>
<td>4.635*</td>
<td>(2.333)</td>
</tr>
</tbody>
</table>

Observations 46
R-squared 0.431
Robust standard errors in parentheses
* p < 0.1, ** p < 0.05, *** p < 0.01

To start with we observe, as expected, that our prediction is more biased where there is a strongest presence of FN voters. The social-media error increases in those areas where the support for the far-right is higher.28

27 Besides, we also tested the bias on MAE in districts with an incumbent candidate but we did not find a significant effect.
28 We tested the same idea with respect to the far-left, using the share of votes won by Mélanchon during the presidential elections to identify the ‘red’ districts. However this aspect seems to yield no bias at all.
Secondly, the MAE is greater when the abstention grows. This could happen because some citizens can easily express their opinion on-line though refusing to cast a vote (perhaps because they feel that their choice will not alter the results or because after all the act of voting is costly: Downs 1957). Social-media analysis then seems less able to provide accurate predictions when voters tend to abstain at a higher rate, while the accuracy should be greater when forecasting elections with a higher turnout. In addition, an increase in the number of eligible voters decreases the error of our prediction given that in low populated districts any misinterpretation of voting behavior ends up being more relevant. Finally we found that the amount of data (tweets) available on-line has a quadratic effect on our dependent variable. Figure 4 in this respect reports the expected value of MAE as the number of tweets change (keeping all the other variables at the values they present in our sample), while Figure 5 displays the corresponding marginal effect of the number of tweets.

As the Figures clearly show, when we are dealing with (relatively) low amount of data, any increase in the number of tweets rapidly reduces the Mean Absolute Error. However, after a certain threshold, a growing number of comments available on Twitter becomes slightly damaging for the accuracy of our prediction, e.g., the expected value of MAE starts to increase once again after having reached a minimum (albeit not returning to
Figure 5: Marginal effect of Number of Tweets on Mean Absolute Error
the large value MAE presented when the number of tweets are comparably very low). A possible explanation has to deal with the idea that larger amounts of tweets contain higher levels of noise that are no longer useful to estimate the actual voting behavior. In addition we know that Twitter can be used by party activists during the campaign. Members of parties that, according to the survey polls, are losing the elections may deliver comments at an higher rate, as an attempt to reduce the gap with the front-runner. These reasons suggest why having more information improves the accuracy of our forecast while having too much information does not necessarily imply the same.29

6 Conclusion

There has been a large discussion about the possibility for the web to become an ‘uncoerced public sphere’, a potential source of direct democracy and mobilization, and how in general the development of social networks more could contribute to increase responsiveness and accountability of real world politics. (Benkler 2006; Gil de Zuniga et al. 2009; Papacharissi 2002; for a critic, refer to: Alvarez and Hall 2011; Hilbert 2009; Hindman 2009; Larssen and Moe 2012). An increasing number of facts, discussed earlier, shows that this is not just an hypothetical scenario. Not surprisingly, in the last years we also assist a growing branch of literature that analyzes social network in order to assess the opinions of internet users and to check whether the attitudes expressed on-line could be used to forecast the voting behavior of the whole population of voters. For all these reasons, being able to rely on techniques apt to measure on-line public opinion becomes a pressing topic.

In this paper we have applied in three (very) different political scenarios a statistical method recently introduced in the literature that performs an automated and supervised sentiment analysis on blogs and social networks, and that improves on traditional Blog Sentiment Analysis yielding more accurate results. From the results of our empirical analyses, we can raise some general claims. First of all, despite internet users are not necessarily representative of the whole population of country’s citizens, our analysis shows, with only few exceptions, a consistent correlation between social-media results and the ones we could obtain from more traditional mass surveys as well as a remarkable ability of social-media on average to forecast electoral results (a so careful prediction that could not be due simply to chance).

This seems to be true for both ‘single issue’ elections (such as Presidential race), in which the preference eventually expressed by an internet user involves only a positive or a negative evaluation among two single options, as well as for more difficult situations to forecast such as the ones in which internet users can choose to express a preference among

This idea is similar to the concept of ‘information overload’, employed by scholars in the field of cognitive psychology to suggest that internet users have to deal with too much information. Such overload can increase the probability of misjudgments when surling the web to find information (Graham 1999; Jha 2007).
a (large) number of different ‘targets’ (such as political leaders or political parties). This is clearly important for the robustness of our results.

Why does this happen? The direction of causality of this pattern (i.e., is the social media opinion the one that tends to become more similar to the average general public opinion as the usage of social networks increases in the large population, or, on the contrary, are social media increasingly driving the general public opinion?) lies beyond the scope of our research. But it is clearly a topic that deserves a further investigation. Besides this main result, the previous analyses also allow to provide some more fine-grained conclusions. For example, sentiment analysis of social network seems to provide more accurate predictions when focusing on the most popular leaders or on the ‘mainstream’ parties. On the contrary, more bias in using social network to make electoral forecasts appears when dealing with radical parties. In particular, supporters of far-right parties tend to be underrepresented on social networks compared to radical left-wing voters. In this respect, the accuracy of predictions based on sentiment analysis could be increased by developing an appropriate set of weights according to the political preferences of social media users, provided this kind of information is available (and reliable). This represents a further topic that deserves to be explored. Secondly, on-line 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. Thirdly, both traditional surveys as well as social networks analyses methods seem to be affected, at least in part, by the same sources of bias, like the strategic behavior and the spiral of silence. On one side, the problem seems to be less severe in the analysis of social networks, where - differently from a traditional survey - the ‘respondents’ take the initiative to express their own opinion and hence a ‘strategic lie’ tends to emerge only when a matter of social acceptability is involved; on the other side, the social network is a forum where people may wish to condition the public opinion, and this may make the communication less spontaneous. This is per se a potentially interesting aspect of the expression of political opinions via the social media; however, were this kind of bias be confirmed by further applications of the methodology, it could be taken into account in the weighting step of the analysis. Finally, and quite interestingly, the possibility of an ‘information overload’ arises as a further factor affecting negatively the analysis of social-media. Summing up, despite the well-known limits and the troubles faced by social-media analysis (Gayo-Avello et al. 2011; Goldstein and Rainey 2010), our results provide reasons to be optimistic about the capability of sentiment analysis to become (if not to be already) a useful supplement of traditional off-line polls.

Acknowledgments

We are grateful to Crimson Hexagon which provided access to the Forsight platform via the Crimson Hexagon Research Grant.
References


Kaufhold, K., Valenzuela, S., and Gil De Zuniga, H. (2010) ‘Citizen journalism and democracy: How user-generated news use relates to political knowledge and participa-


