Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' policy preferences. An application to Italy and France

Keywords: *sentiment analysis, text mining, text analytics, social media, political forecast*

The exponential growth of social media and social network sites, like Facebook and Twitter, raises the possibility to delve into the web to explore and track the (policy) preferences of citizens. 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 "sentiment analysis" (SA) 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 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. 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 "freedom from irrelevant alternatives" condition (e.g., including the German 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 foresight (Chung and Mustafaraj, 2011). Accordingly, other studies tried to improve this stream of research by means of sentiment analysis.[1] 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) show 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 as well as polls in predicting the results of 2011 Dutch Senate Election (Tjong Kim Sang and Bos, 2012), while 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 in Hopkins and King (2010) (HK, from now on). As we will discuss, this method presents various advantages compared to traditional SA techniques. We will employ it in three 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 actual electoral results.

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

1. How to scrutinize citizens' preferences through social-media: a novel method (and some pros and cons)

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 (approximately 2.5 billion people).[2] Among them, 72% of the internet population is active on at least one social network,[3] like Facebook (over 800 million of users, 12% of the world population)[4] or Twitter (140 million of active users).[5]

This is by itself a good reason to recall the importance of analyzing social media, albeit by far not the only one. Recently, social network sites have started to wield substantive effects on real world politics: they have been used to organize demonstrations and revolts, for instance during the 'Arab spring' (Ghannam, 2011; Cottle 2011)[6]; to engage individuals in mobilizations (Segerberg and Bennett, 2011; Bennett and Segerberg, 2011); to build social movements and political parties, like the Pirate Party in Sweden and Germany or the Italian '*Movimento 5 Stelle'*, which use the web to set the party line and to select candidates.[7] Social media are also often adopted 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 explicitly invited Italian citizens to provide suggestions through web on two topics: the spending review[8] and the legal value of the university degree.[9]

These examples provide a contribution to the discussion about the possibility for the web to become an "uncoerced public sphere". So far, several authors (e.g.: Benkler, 2006; Downey and Fenton, 2003; Langman, 2005; Papacharissi, 2002) have debated about the potential of the new media to act as a 'habermasian public sphere'. While some of them suggest that internet and the social media are potential sources of direct democracy, which may contribute to increase responsiveness and accountability of real world politics (Benkler, 2006; Gil de Zuniga *et al.*, 2009; Papacharissi, 2002), other authors proposed diverging views that criticized the idea of the web as a public sphere (Alvarez and Hall, 2011; Hilbert, 2009; Hindman, 2009; Larsson and Moe, 2012).

Notwithstanding this debate, 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*.[10] 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 in Hopkins and King (2010).

The main advantage of the HK method is that it performs a supervised and automated sentiment analysis. The traditional approach to sentiment analysis is in fact based on the use of ontological dictionaries: this means that 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). The strong drawback, on the other side, is the difficulty in classifying opinions expressed through ironic or paradoxical sentences, or in appreciating all the language nuances (specific jargons, neologisms, etc.): 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 is based on a two-stage process. The first step involves human coders and consists in 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, in the second stage. Human coders are of course more effective and careful than ontological dictionaries in recognizing all the previously discussed language specificity and the author's attitude to the subject (Hopkins and King, 2012). Moreover, human coding is better suited to identify the (ever-present) problem of spamming in social communication. This is of course important, given that spamming can have an impact on the accuracy of the final result. At the second stage, the automated statistical analysis provided by the HK algorithm extends such accuracy to the whole population of posts, allowing for properly catching the opinions expressed on the web. Indeed, the expected error of the estimate is around 3%.[11]

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, posts) in the data set are decomposed into their own single words: consequently, each unit is represented by the vector of the terms used, which we call "word profile" of the unit.[12]

The formal background of the method is simple (for further details see: Hopkins and King, 2010). Indicate by **S** the word profiles 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:

(*) $P(\mathbf{S}) = P(\mathbf{S}|\mathbf{D})\, P(\mathbf{D})$, 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 $P_T(\mathbf{S}|\mathbf{D})$, 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:

$$P_T(\mathbf{S}|\mathbf{D}) = P(\mathbf{S}|\mathbf{D})$$

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

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.

Besides, while classic web analyses allow only distinguishing between positive and negative references to a particular topic, the HK enables to measure the intensity of preferences (i.e., how much pro or con). In addition it can also allow catching the opinions expressed by the net, which are not necessarily connected to a positive or a negative statement (and hence cannot be captured on an ordinal scale): as an example, the opinion on a new party can be "I like that party's leader, but his party is too radical for me", or "Their manifesto is just perfect to me. The problem is that party will never pass the electoral threshold". The human-coding phase can identify all these categories; the subsequent automatic classification estimates their relative weight in the overall opinion distribution.

Broadly speaking, there are several social-media that could be analyzed. Here 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 in 2009. When we come to the countries analyzed in this work, we observe that 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, approximately 20% of the Italian population (Mazzoleni et al., 2011). Among them one out of three (6% of the whole population) released political comments and opinions.[13] 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.

To download the data employed in the present paper we have 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/).[14] For the Italian case the analysis of the tweets have been performed using directly the ForSight platform provided by *Crimson Hexagon*, while for the French two cases the data have been collected through *Voices from the Blogs* and the analysis have been then run in R.[15]

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 SA, 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 SA 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, we can argue that Internet may represents to a large extent an arena where the public expression of preferences is uncoerced by any established political agenda because users are free to express themselves (Savigny, 2002).[16] In this vein, social network should be in the position to be less affected by the spiral of silence (Noelle-Neumann, 1974),[17] 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 socio-economic 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 show that senior citizens are underrepresented on the web (Fox, 2010) and there is a prevalence of highly educated male individuals (Wei and Hindman, 2011).[18] 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).[19] Along this vein, Best and Krueger (2005: 204, italic 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*'. 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) (source: ISPO, Istituto per gli Studi sulla Opinione Pubblica). 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 media 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 media analysis compared to traditional off-line surveys. 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 apply. 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" as an all) the preferences of a wider audience: as a result, it could happen that the preferences expressed through social media today will affect (predict) the opinion of the whole population tomorrow (O'Connor et al., 2010).

In the next 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 distinguished political phenomena: leaders popularity, Presidential and Legislative national elections.

2. Comparing Italian leaders' popularity ratings in 2011

The 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 13 January 2011 to 20 October 2011: they are based on a sample of around 800 respondents 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.[20]

Similarly, the popularity of each leader from social media has been estimated as the percentage of his positive posts over the sum of his 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 ranges: 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 the 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 more than five 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 visible Italian leaders during 2011) and on the other all 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

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) of the predictions on leaders popularity as they arise from social-media as compared to the scores obtained from Mass-surveys. The MAE has been indeed 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). This is what we do here as well. As can be seen, despite being quite relevant at the beginning of 2011 (around 13 points), this absolute difference tends to markedly decrease as times goes by. [21]

Figure 1

To sum up, albeit being just an exploratory analysis, our results provide some quite interesting insights:

- At least for the most visible leaders, the two measures of popularity 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.

- 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 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 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.[22] 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: see Figure 1). 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 good news for the electoral forecasting ability of social-media analysis. The next two sections are devoted to explore precisely this latter possibility.

3. Electoral campaign and social media (1): the 2012 French presidential 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 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.[23] 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 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

First of all, on 28 April 2012 we found a peak in favor of Hollande. In those days Sarkozy was dealing with a document that 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 media 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 had an echo on-line and contributed to a 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 turn blamed the media 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 1 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 2 May 2012. 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 2 May 2012, two third were comments to the debate). In line with other analyses our estimates confirm that Hollande has prevailed during the debate.[24] 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 illustrated how 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 us to correctly forecast the outcome of the polls, wielding predictions that are very close to those made by traditional surveys (see also: Larsson and Moe, 2012; Tjong Kim Sang and Bos, 2012).

4. 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 a 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.[25]

That noted, 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

The data on the French Legislative election allow 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, Djion, 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.[26] 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 MAE of our prediction. We have estimated three different models. In the first one, we include 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 add 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 is 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 is running to seek the re-election. Finally, in Model 3 we add the 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 do actually cast a vote. Data have been analyzed through OLS regression. Table 2 reports the results.

Table 2

From Model 1 we observe that any growth of the information available on-line improves our predictive skills. For instance, an increase of 1,000 tweets analyzed lowers our error roughly by a quarter point. Conversely, the MAE is greater when the *Abstention* growths (the same concerns affects traditional off-line pre-electoral polls: see Crespi, 1988). 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 (a 10% increase in *Abstention* raises 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 some control variables (Model 2). Our prediction does not appear therefore biased by the stronger presence of FN or left-wing voters. While we know that incumbent candidates usually benefit from an advantage when seeking re-election (Ansolabehere and Snyder, 2002; Gelman and King, 1990), it has also been argued that elections are 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 skills.

Finally, in Model 3, we tested the conditional effect of *Number of Tweets* and *Abstention*. The coefficient of the interaction term is 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 intentions, and such effect enlarges as turnout growths. 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 would express themselves on Twitter instead of casting a real vote, the Mean Absolute Error tends to increase. This (original) finding is quite interesting given that it clearly shows the strict relationship incurring between what happens on-line and off-line in terms of our ability to extract reliable measures from social-media analysis.

Figure 4

Conclusions

The growing usage of social media by internet users, who express their opinions, through social network sites, on a wide variety of topics including their political and policy preferences, has raised the interest about the opportunity to exploit this information to better understand the link between political preferences and political behavior. Not surprisingly, in the last years we assist a dramatically increase in the number of works that analyzes social media in order to assess the opinions of internet users and to check whether the attitudes expressed on-line can be eventually 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 Sentiment Analysis wielding 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 to forecast on average electoral results (a so careful prediction that could not be due simple 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 a (large) number of different *targets* (such as political leaders or political parties). This – together with the fact that the political scenarios here analyzed come from 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.

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 as an all increasingly driving (or anticipating) 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 media seems to provide more accurate predictions when focusing on the most popular leaders or on *mainstream* parties. On the other side, the accuracy of predictions based on sentiment analysis 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 underrepresented on social networks compared to radical left-wing voters), 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.

Finally, some of the potential bias that arise from social-media analysis may be softened in the long run, as the usage of social network increases: as we have shown, when a growing number of citizens' express on-line their opinion and/or voting choice, the accuracy of social media analysis increases, provided internet-users act consistently on that (by, for example, confirming their (declared) on-line preference by casting a (real) vote).

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.

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: 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, W. Lance and Alexandra Segerberg (2011). Digital media and the personalization of collective action: Social technology and the organization of protests against the global economic crisis, *Information Communication and 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 review* 6.

Chadwick A (2008) Web 2.0: New challenges for the study of e-democracy in an era of informational exuberance. *I/S: Journal of Law and Policy for the Information Society* 4(3): 9–42

Chung J and Mustafaraj E (2011) Can collective sentiment expressed on twitter predict political elections? Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, USA.

Cottle S (2011) Media and the Arab uprisings of 2011: Research notes. *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.

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 and Row.

Falck O, Gold R and Heblich S (2012) E-Lections: Voting Behavior and the Internet. *IZA Discussion Paper*.

Fox S (2010) Four in ten seniors go online. Pew Research Center, URL: http://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* DOI: 10.1080/19331681.2012.705080

Freeman SF and Bleifuss J (2006) *Was the 2004 Presidential Election Stolen?: Exit Polls, Election Fraud, and the Official Count*. Seven Stories.

Gayo-Avello D (2011) A warning against converting social media into the next literary digest. *CACM*.

Gayo-Avello D, Metaxas P and Mustafaraj E (2011) Limits of electoral predictions using social media data. Proceedings of the International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

Gelman A and King G (1990) Estimating Incumbency Advantage Without Bias. *American Journal of Political Science* 34: 1142–1164.

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

Gibson RK, Lusoli W and Ward S (2008) Nationalizing and normalizing the local? A comparative analysis of online candidate campaigning in Australia and Britain. *Journal of Information Technology & Politics* 4(4): 15–30.

Gil De Zúñiga H, Puig-I-Abril E and Rojas H (2009) Weblogs, traditional sources online and political participation: An assessment of how the internet is changing the political environment. *New Media & Society* 11(4): 553–574.

Gloor PA, Krauss J, Nann S, Fischbach K and Schoder D (2009) Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. *CSE* 4. 2009 International Conference on Computational Science and Engineering: 215–222.

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

Hilbert M (2009) The maturing concept of e-democracy: From e-voting and online consultations to democratic value out of jumbled online chatter. *Journal of Information Technology & Politics* 6(2): 87–110.

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.

Huber J and Hauser F (2005) Systematic mispricing in experimental markets evidence from political stock markets. In 10th Annual Workshop on Economic Heterogeneous Interacting Agents. Essex, UK.

Jansen HJ and Koop R (2005) Pundits, Ideologues, and Ranters: The British Columbia Election Online. *Canadian Journal of Communication* 30(4): 613–632.

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

Jugherr A, Jürgens P and Schoen H (2011) Why the pirate party won the German election of 2009 or the trouble with predictions: A response to Tumasjan A, Sprenger TO, Sander PG and Welpe IM ‘‘Predicting elections with twitter: What 140 characters reveal about political sentiment’’. *Social Science Computer Review*.

Kaufhold K, Valenzuela S and Gil De Zuniga H (2010) Citizen journalism and democracy: How user-generated news use relates to political knowledge and participation. *Journalism & Mass Communication Quarterly* 87(3–4): 515–529.

King G, Pan J and Roberts M (2012) How Censorship in China Allows Government Criticism but Silences Collective Expression. Available at: http://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. URL: http://languagewrong.tumblr.com/post/55722687/predicting-polls-with-lexicon.

Madge C, Meek J, Wellens J and Hooley T (2009) Facebook, social integration and informal learning at university: It is more for socialising and talking to friends about work than for actually doing work. *Learning, Media and Technology* 34(2): 141–155.

Marangoni F (2011) The Premature End of the Fourth Berlusconi Government: An Updating of Indicators of the Government's Legislative Performance and Productivity. *Bulletin of Italian Politics* 3(2): 357–369.

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

Metaxas PT and Mustafaraj E (2010) From obscurity to prominence in minutes: political speech and real-time search. *WebSci10: Extending the Frontiers of Society OnLine*. Available at: http://bit.ly/h3Mfld.

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

Mustafaraj E, Finn S, Whitlock C and Metaxas PT (2011) Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail. *Proceedings of SocialCom/PASSAT*: 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: 43–51.

O'Connor B, Balasubramanyan R, Routledge BR and Smith NA (2010) From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

Östman J (2012) Information, expression, participation: How involvement in user-generated content relates to democratic engagement among young people. *New Media & Society* 14(6): 1004–1021.

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: Princeton University Press

Savigny H (2002) Public Opinion, Political Communication and the Internet. *Politics*, 22(1): 1–8.

Segerberg, Alexandra and W. Lance Bennett (2011). Social media and the organization of collective action: Using twitter to explore the ecologies of two climate change protests, *Communication Review*, 14(3): 197-215

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

Sunstein CR (2002) The Law of Group Polarization. *Journal of Political Philosophy* 10(2): 175–195

Tjong Kim Sang E and Bos J (2012) Predicting the 2011 Dutch Senate Election Results with Twitter. Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, Avignon, France, 2012.

Tumasjan A, Sprenger TO, Philipp GS and Welpe IM (2010) Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

Upton G Jr (2010) Does Attractiveness of Candidates Affect Election Outcomes?. Available at: http://com/lib/files/AttractivePoliticians.pdf.

Véronis J (2007) Citations dans la presse et résultats du premier tour de la présidentielle 2007. Available at: http://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. Annual Meeting of the American Political Science Association, 1-17. Boston, MA.

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 and Cao L (2010) The Wisdom of Social Multimedia: Using Flickr For Prediction and Forecast. *ACM Multimedia 2010 International Conference*. Key: citeulike:7968236.

Tables

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.*

	avg. difference	*st.dev.*	*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

Table 2. *OLS regression of Mean Absolute Error*

	(1)	*(2)*	*(3)*
VARIABLES			
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 X 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

Figures

Figure 1. *The MAE of leaders' popularity ratings between Mass-Surveys and social media estimated over time*

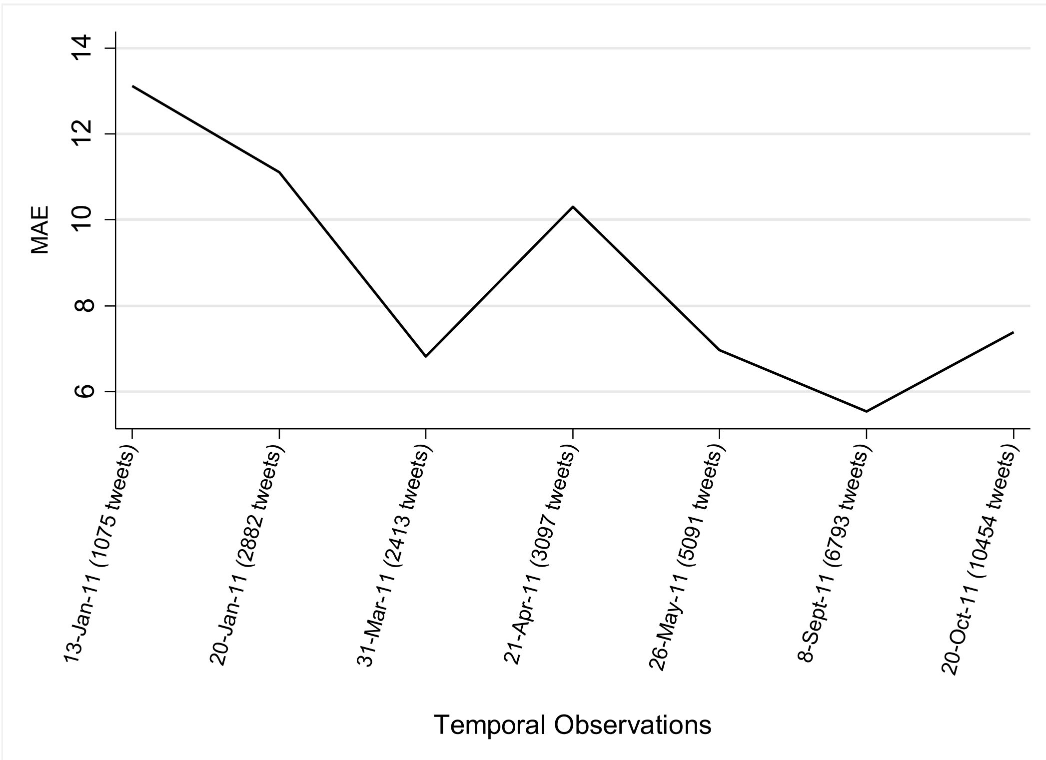


Figure 2. *Flow of preferences expressed on Twitter during the electoral campaign for the second round of 2012 French Presidential election*

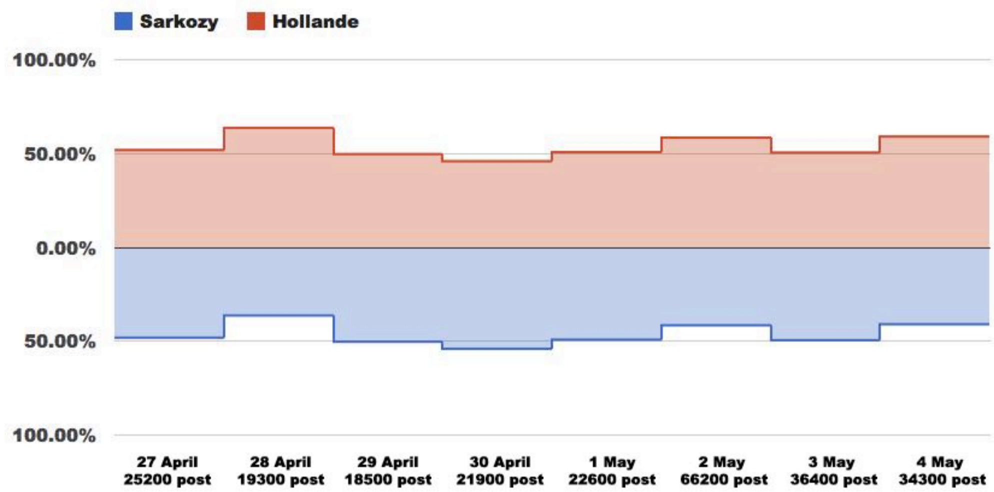


Figure 3. *Predicted and actual vote shares related to the first round of the 2012 French Legislative elections*

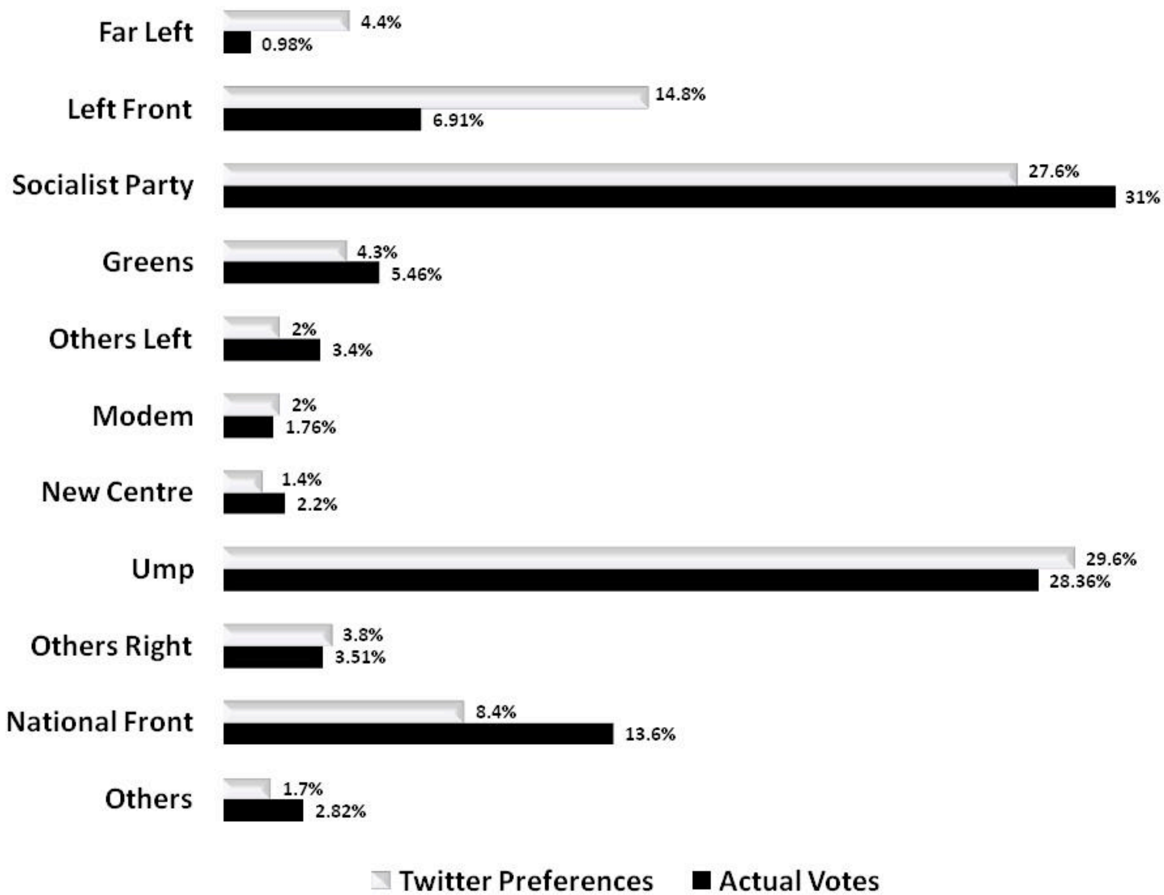
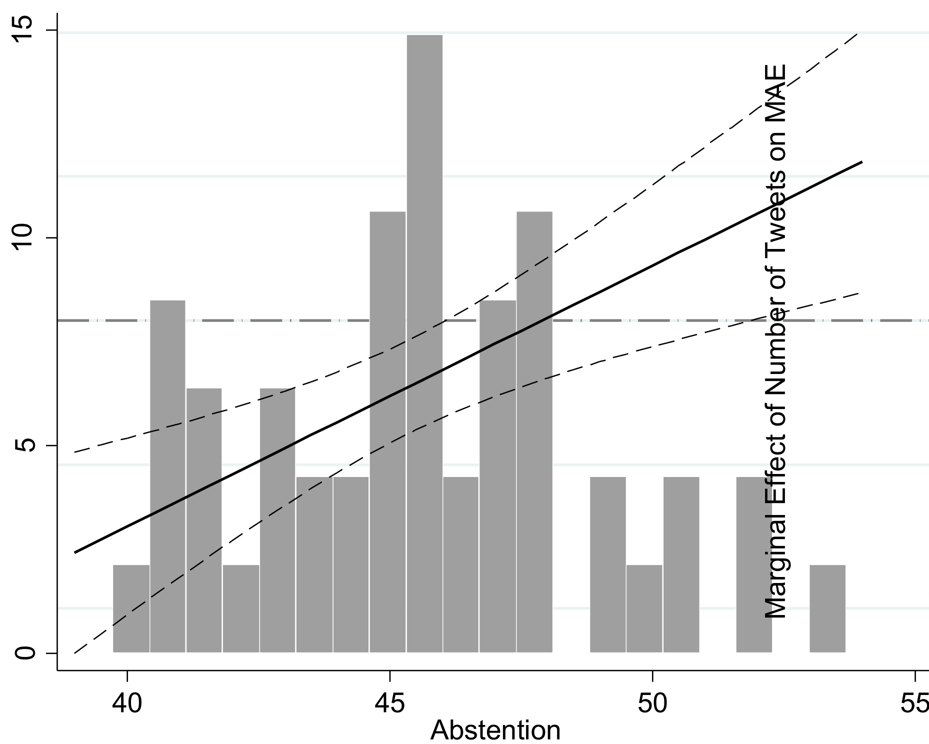


Figure 4. *Marginal effect of the Number of Tweets on the Mean Absolute Error (with 90% confidence interval)*



Notes

[1] Sentiment analysis consists in analyzing texts to extract information.

[2] Source: International Telecommunications Unions (http://www.itu.int/ITUD/ict/facts/2011/material/ICTFactsFigures2011.pdf).

[3] Source: Social Network around the World 2010 (http://www.slideshare.net/stevenvanbelleghem/social-networks-around-the-world-2010).

[4] Source: http://www.internetworldstats.com/facebook.htm.

[5] Source: 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 on 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 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. In Italy, the *Movimento 5 Stelle* also reported surprising results during local elections held between 2009 and 2012, up to a point that survey polls published during September 2012 attribute to that party a striking 20% in the voting intention of Italian voters.

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

[9]http://archiviostorico.corriere.it/2012/marzo/03/Laurea_via_alla_consultazione_online_co_9_120303011.shtml.

[10] 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 have affected

the content of the electoral campaigning and the political communication by 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).

[11] http://www.crimsonhexagon.com/quantitative-analysis/. From our replications, the root mean square error of the estimates drops until 1.5% when the number of hand-coded documents increases up to 500.

[12] In other terms, a "word profile" is a vector made of 0's and 1's: 0 when a term does not appear in the unit (but it is used in some other units) and 1 when a term appears in the unit.

[13] Source: Osservatorio Politico ISPO (Istituto per gli Studi sulla Pubblica Opinione). We thank ISPO for having shared the data.

[14] The population of tweets collected consists of all the tweets posted during the temporal-period considered (see below) that 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 analysis in both Italy and France).

[15] Script and data are available upon request.

[16] 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 *et al.*, 2012).

[17] 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.

[18] 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 et al., 2011;

Tumasjan et al., 2010). In addition, some accounts are fake (Metaxas and Mustafaraj, 2010). Finally, and by definition, we can observe on-line only the opinion of those who have decided to express their attitudes (Gayo-Avello et al., 2011).

[19] Gender constitutes a partial exception given that, despite equal participation to social networks, males tend to express their political views more than females do.

[20] The survey question on which the popularity rating is based is the following one: “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: 0 meaning completely negative judgment, 10 completely positive judgment and 6 sufficiently positive”.

[21] The peak in MAE registered in April 2011 corresponds to a situation in which the information on leaders’ popularity from the Mass-Surveys covers just two Italian leaders out of eight (i.e., Berlusconi and Fini). If we eliminate this eccentric temporal observation, the average value of MAE over the period is less than eight points.

[22] The Berlusconi IV cabinet was weakened by the split of the *formateur* party, People of Freedom (PDL), in 2010 (Ceron, 2011). 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 (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.

[23] For instance, Ipsos predicted his victory with 52.5% and TNS-Sofres with 53.5%.

[24] See http://tempsreel.nouvelobs.com/election-presidentielle-2012/20120502.OBS9654/sarkozy-ou-hollande-qui-gagne-le-debat-l-analyse-sur-twitter.html.

[25] 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] We excluded Paris due to its broad size that makes it harder to establish a link between the origin of each post and the electoral districts existing in the city.

Response to: “Every Tweet Counts? How sentiment analysis can improve our knowledge of citizens’ policy preferences” by Stefano Iacus, University of Milan

---

presented at:
Social Media and Political Participation
La Pietra, NYU Firenze

discussant:
Richard Bonneau, NYU

Summary of method

- Method for sentiment analysis built based on method presented by Hopkins and King, 2010

$$\underset{2^K \times 1}{P(\boldsymbol{S})} = \underset{2^K \times J}{P(\boldsymbol{S} \mid D)} \; \underset{J \times 1}{P(D)}$$

- Key strength of paper is careful comparison to surveys and voting results in 3 contexts: Italian Elections, French National election, French Legislation

- Use of geotagging/location to assign tweets to region

Main results

- Good predictive performance for highly visible candidates in Italian election, but
  R between predicted sentiment (for Italian election) ranged from 0.05 to 0.933

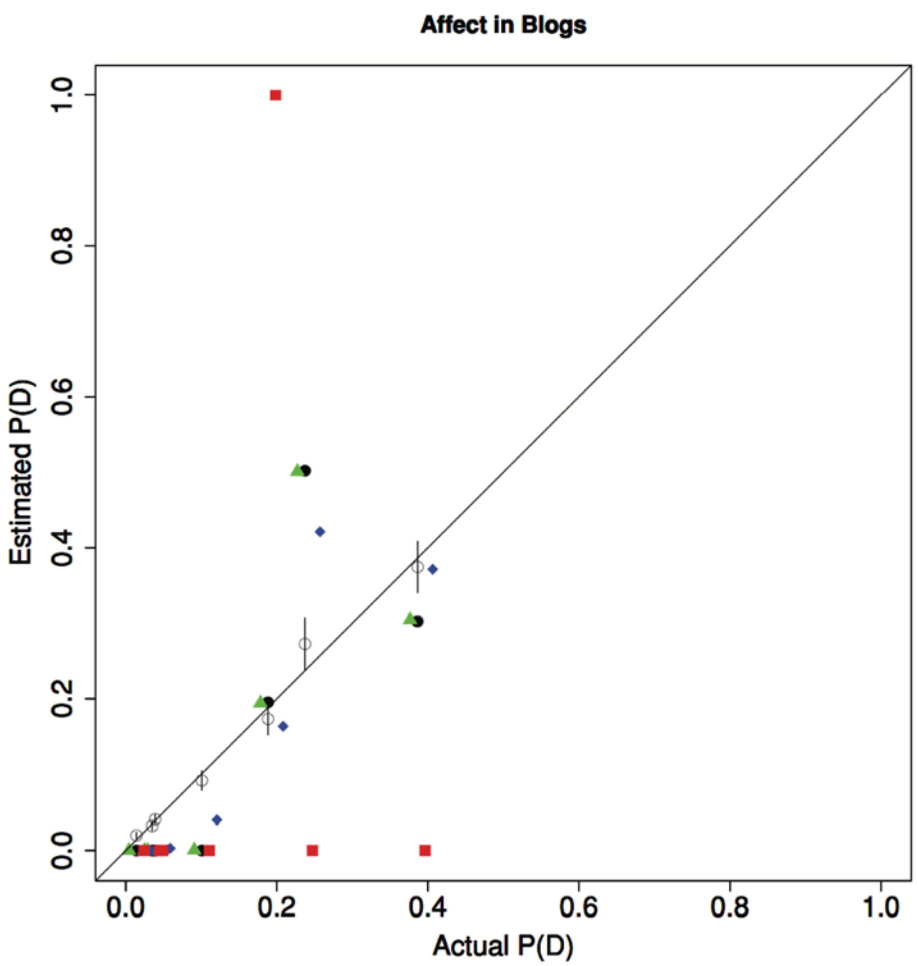
- Results better for French elections. Some bias suspected based on difference in conservative and liberal parties.

- number of tweets and fraction Abstention were predictive of sentiment analysis MSE

- some successes and some places where we need to improve the method

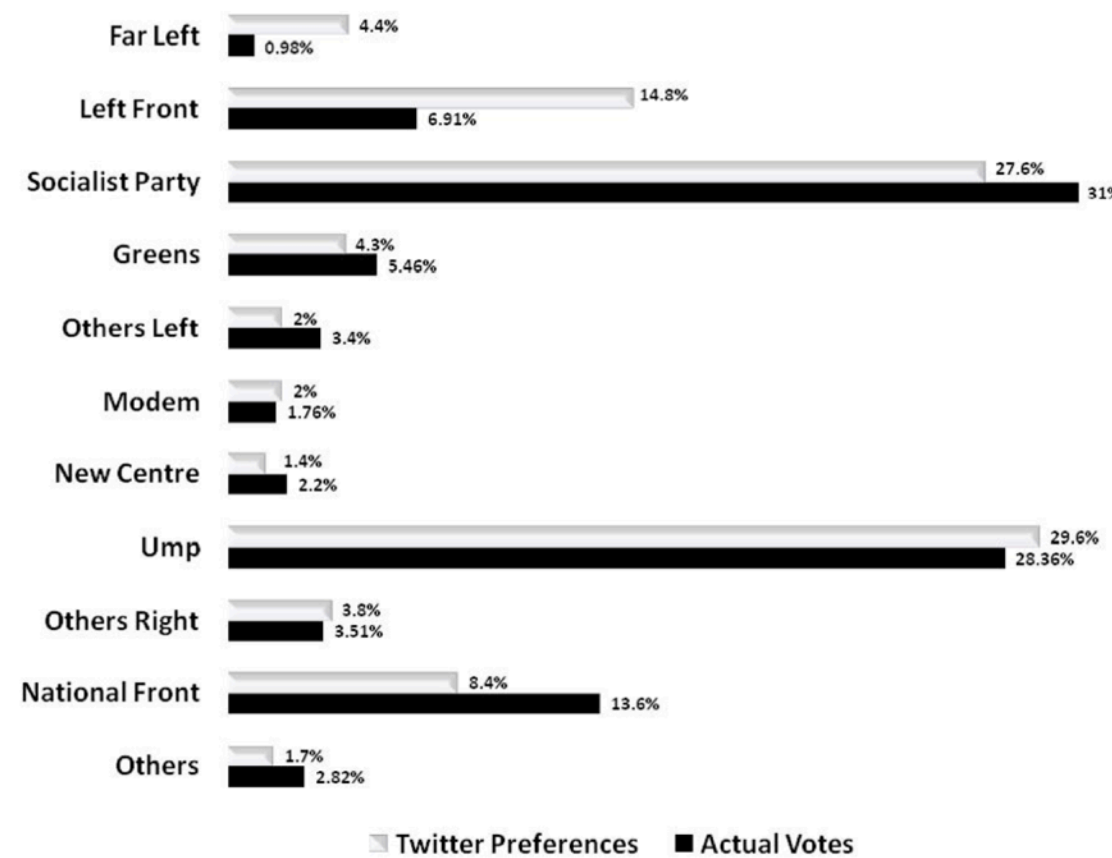
HK method continues to work well in new places

FIGURE 3 Out-of-Sample Validation



(Hopkins & King 2010)

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



(Iacus, 2013 (today))

methods details I wanted

- # of tweets coded and some analysis of how performance vs. # number coded differed for different parties

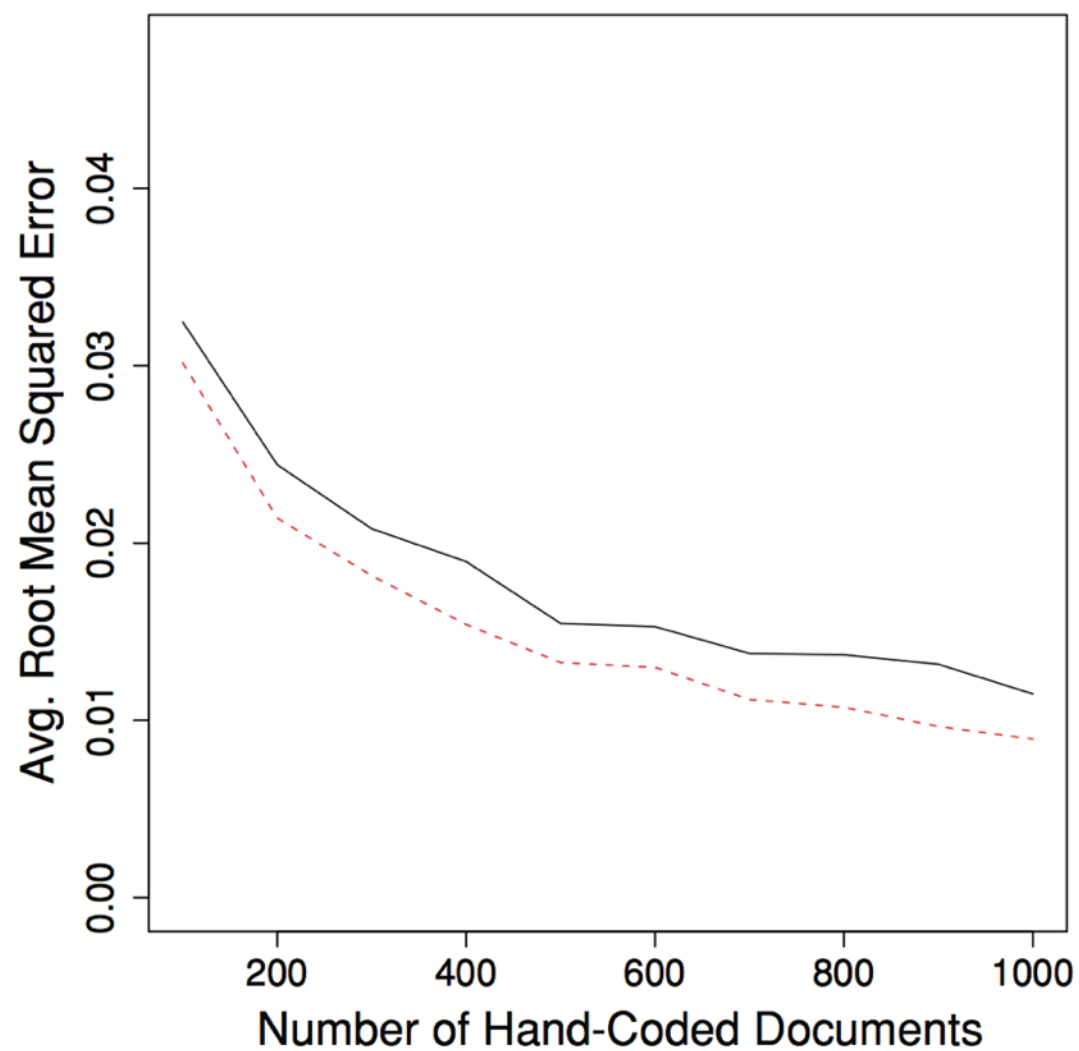
- # of words in your subset used in HK method

- Examples of mass survey and coded tweets

- discussion of how good these methods need to be operationally useful to political operative, useful to those developing theory, etc.

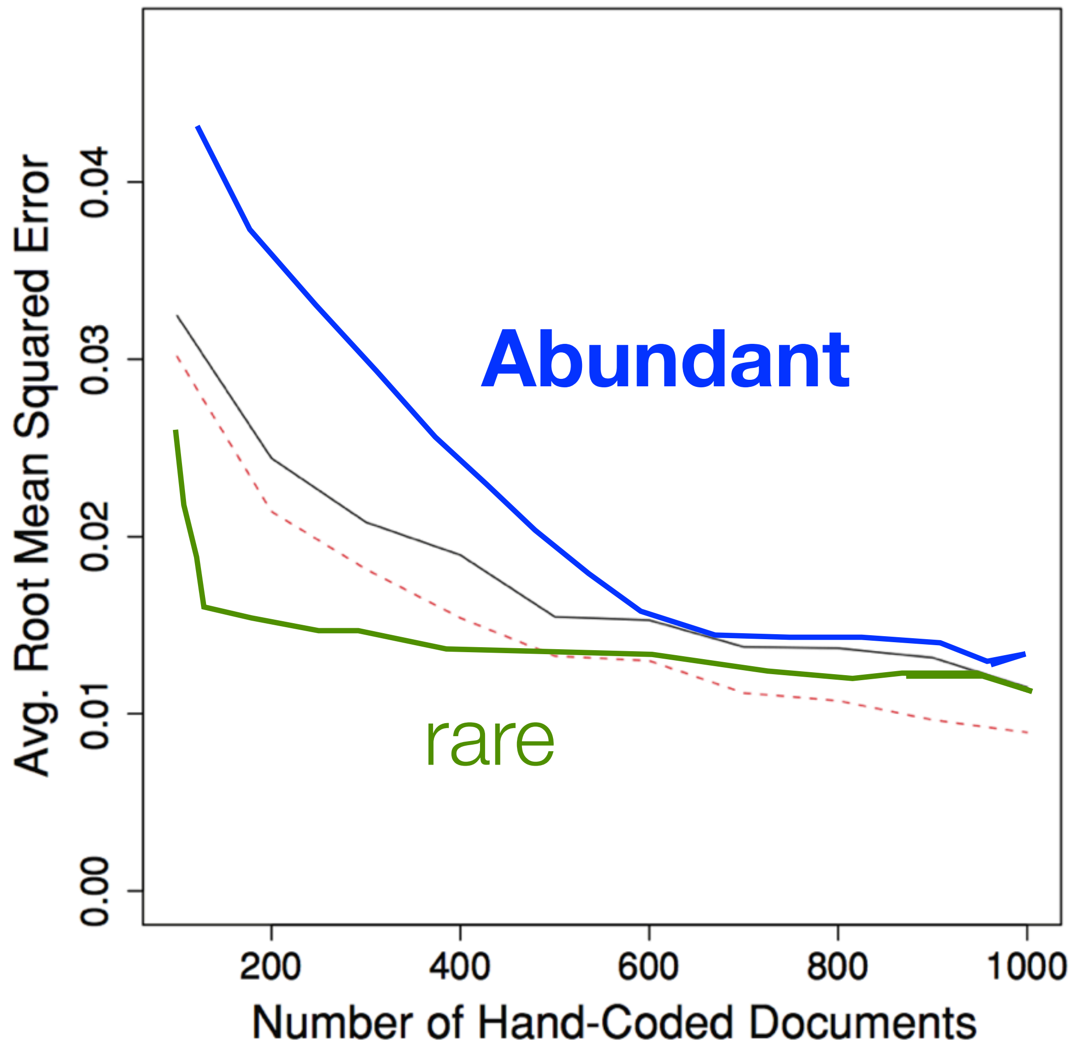
how much coding is enough: split by label abundance?

FIGURE 5 Average Root Mean Square Error by Number of Hand-Coded Documents



how much coding is enough: split by label abundance?

Figure 5 **Average Root Mean Square Error by Number of Hand-Coded Documents**



Could the network help get from tweet -> **person**