

Do social media affect public discourses? A sentiment analysis of political tweets during the French Presidential Election campaign

Leendert de Voogd (TNS, Political & Social, leendert.devoogd@tnsglobal.com)

Pascal Chelala (TNS opinion, Brussels, pascal.chelala@tns-opinion.com)

Steve Schwarzer (TNS opinion, Brussels, Steve.schwarzer@tns-opinion.com)

Paper to be presented at the AAPOR 67th General Conference, Orlando, Florida, May 2012

-- First Draft, PLEASE DO NOT QUOTE OR CITE WITHOUT PERMISSION FROM ONE OF THE AUTHORS --

Abstract

This paper analyses the relationships that may exist between political polls, traditional media exposure and social media content during the French Presidential Campaign in 2012. It focuses more specifically on the sentiment shared spontaneously by tweeterers in their messages about François Hollande and Nicolas Sarkozy and aims at demonstrating how citizens and supporters utilize Twitter strategically as a new form of militantism and agenda setting.

On the one hand, microblogging, and especially Twitter does allow for an easy-to-access and easy-to-use platform to exchange ideas and opinions and might therefore also intensify connections between different individuals. On the other hand, these social media platforms offer interesting opportunities for researcher to focus on what the citizens express and share spontaneously and without any filter from a totally different angle. The Internet's role in the public sphere and social capital mostly discusses the possibilities of the Internet to cultivate social contact, mobilize political opinion, socialize and educate people, enable people to share their opinions to a wider public (one2many communication), and enhance the contact between citizens and their political representatives. We connect measures of media exposure, public opinion measured in a daily poll over three month with sentiment measures of the content of tweets. We find a high correlation between all three domains, which indicates that the analysis of social media content can help us to measure public salience.

Introduction

The exponential growth of Twitter has started to draw the attention of researchers from various disciplines. There are several streams of research investigating the role of Twitter in social media, product marketing, and project management. While findings have provided us with a general understanding of why and how people use microblogging services, they have not yet explored the use of this new communication device in political discussions online.

Social network sites are a popular form of social media, but in the end all the different applications share one commonality: the user generated content that is open and available to others (Kushin and Yamamoto 2010, p. 612; Bruns 2011). Online expression is functionally distinct from simply consuming content online. Social media allow users to not only seek information but also interact with others through online expression such as posting political commentaries on blogs and social network sites and sharing other online or multimedia content (Kushin and Yamamoto 2010, p. 613). Boyd states that social media is about the collective action and is organized around human interaction (Boyd 2007; Boyd and Ellison 2007). The communication via Twitter is not only open for discussion but, by its nature, Twitter is also directed to the global public by default (Bruns 2011, p. 2). Twitter publics can be understood as interconnected and fluid conversations constituted by the interplay of users connecting and relating to each other. The sphere based on this shows several connection to other media content, while Twitter is also integrated in other media content.

In these or other ways, social media such as blogs, Tweets, wikis, and social networks are all about speeding up and enriching communication. They go beyond the old mode of one to one communication and enable communication from one to many via a blog post or a tweet or from many to many as on one's profile. People are able to create their own audience, which can become big and even bigger (Hawn 2009).

Media and other communication technologies are not mere instruments that can be entirely separated from culture, society and values. They change the way how we do things, how we live, what we do and maybe even how we are. Moreover, media techniques mediate our relation to the world, as they strongly influence the way we interpret the world, with whom we interact and how we perceive what is going on outside (Verbeek 2005; Coeckelbergh 2010). While studying public opinion and the effects of media and social media on humans and society, we need to be aware that the medium also explains how a message is conveyed. New technologies change our perceptions, change our world and also change our communication and information behaviours.

The social medium Twitter provides a unique opportunity for agenda setting and for researchers a new source for measuring public opinion. Even if the sample on life experiments in the social media environment cannot be structured, one can think about the data collection as an indirect measurement of public salience, which is normally measured via public opinion polling. Research on agenda setting was long restricted to the cross-lagged relationship between mainstream media and this is reflected in the wider public (Severin and Tankard 2001). Especially the salience on political issues is linked to public opinion polls, even if they are a large investment in time and costs, these kinds of surveys are built on significant and representative samples. But social networking and microblogging platforms

might provide an additional measurement to capture the expression of a more spontaneous type of public opinion. Nowadays, these kinds of services are widely used and users are more open to share even very private information with a user-defined public. They share opinions, news, thoughts and musings and at least on Twitter, this information is (in the vast majority of cases) available to everyone. This means that a huge discourse arena opens up (Brustein 2010, cited after Vargo 2011). For research, Twitter is a valuable source of information as virtually all Tweets on the website are searchable and quantifiable (Java, Song et al. 2007).

This study tries to first see, whether there is a link between media exposure for candidates and mentions on the social media platform twitter. As Larsson and Moe showed for the 2010 Swedish election, the main spikes of activities were linked "either televised debates, or the media coverage of offline events such as political rallies" (2011, p.19). User-generated content can be seen as a collection of a broad information exchange, in which users not only post their own opinions, but reflect on and discuss the comments of others and external sources (e.g., the mass media).

Our study investigates whether one can make use of social media not as an alternative but an additional source to understand the formation of public opinion in our societies. The focus of our study is on the comparison of the information extracted from social media with accepted benchmarks such as public opinion polls and quantitative media content analysis.

Classical media and microblogging - exchange of information online & offline

Twitter sentiment analysis has not yet been applied to research regarding the political debate online. While several scholars have debated the potential of weblogs as a forum for democratic debate, "empirical research on deliberative democracy has lagged significantly behind theory" (Delli Carpini, Cook et al. 2004, p. 316). A few researchers have empirically examined Internet discussion boards as a vehicle for political deliberation (Jansen and Koop 2005). The latter have defined the exchange of substantive issues as an indicator of deliberation and the equality of participation as a measure of the deliberative quality of blog-based discussion (Jansen and Koop 2005). While they have found discussion boards and blogs to be dominated by a relatively small number of users, it is unclear whether these findings also apply to the political debate on microblogging website, such as Twitter.

Recent scholarly work on political blogs has focused on their effect on real-world politics, such as complementing the watchdog function of the mainstream media and mobilizing supporters, but largely ignored the reflection of off-line politics in the digitally enhanced public sphere. However, there are only a few studies exploring the reflection of the political landscape in "traditional" weblogs and social media sites. For instance, Williams and Gulati (2008) have found that the number of Facebook supporters can be considered a valid indicator of electoral success. But, it is not quite clear yet, whether this can be easily transferred to Twitter as tweets are much shorter than a traditional blog.

Thus, a basic question is whether 140-character messages can contain differentiated information regarding the electorate's political preferences. Preliminary results from two recent reports suggest that microblogging content may be a good predictor of election results. A conference paper analysing

the correlation between candidate mentions on Twitter and the results of the Japanese national election reports that in more than 80% of all constituencies the most mentioned candidate won the election (Suenami & Yutaka, 2010 cited after Tumasjan, Sprenger et al. 2010). A similar survey of candidate mentions on Twitter during the 2010 U.K. election, presented by a website that aggregates political tweets, finds the predictions of the national share of vote to be better than most opinion polls with an average error of only 1.75 percentage points (Tweetminster 2010). Also Jungherr et all 2011 and Tumasjan et all 2010 were able to link online content to election outcomes and whereas the latter concludes that the “The mere number of tweets reflects voters’ preferences and comes close to traditional election polls, while the sentiment of political twitter messages closely corresponds to the electorate’s sentiment” (2010, p. 13), Jungherr et all are much more pessimistic in saying that “The number of party mentions in the Twittersphere is thus not a valid indicator of offline political sentiment or even of future election outcomes” (2011, p. 5).

As the empirical evidence is not clear with respect to the possible use of microblogging content to draw conclusions on the offline world, there is still some evidence that encourages us to believe that the information stream on Twitter can be aggregated in a meaningful fashion in order to make accurate, albeit not necessarily representative, predictions. We therefore believe that we can leverage Twitter as a new information market, also opening up fields for enhanced qualitative research based on spontaneous feedback.

Political deliberation online

While the online forum is open by its nature, the communication on this kind of websites is individually structured. Users follow their personal awareness streams (Naaman, Boase et al. 2010; Naaman, Becker et al. 2011). Because studies suggest that people primarily use social network sites to keep in contact with their existing groups of friends and acquaintances and that they use these specific sites to learn more about individuals they meet offline (Lampe, Ellison et al. 2006; Boyd and Ellison 2007). Social networking sites may increase both bonding and bridging social capital, because social network sites boost the sense of community, which includes both the bridging (to other users) and bonding (intensifying relationships with known users) social capital(Putnam 2000). Of course, also in social network users are significantly more likely to connect with someone they know off-line (Lampe, Ellison et al. 2006). But, of course, social network sites also allow users to join groups and causes that could potentially bring them in contact with a diverse group of people.¹

Twitter as a deliberation forum

In comparison with other social media portals, Twitter is recognised as a non-private medium and the conversation is not limited to person to person or person to friends communication. While conservation happen in the public between different networks of people, the information retrieved from sides such as Twitter can be seen as indirect measurements of topics that are salient to the general public (Vargo

¹ Nearly 80% of Facebook and MySpace users in one survey had joined a group Royal, C. (2008). User-generated content: How social networking translates to social capital. Annual meeting of the Association for Education in Journalism and Mass Communication, Chicago, IL., suggesting that social network sites can also create bridging capital.

2011). Arguing with McLuhan (1964), the content of a medium is always built on the content of another medium and cannot be seen as something totally independent from the means which are already available to people. With respect to the total amount of information available to people, it is even more difficult for users to differentiate between important and unimportant messages, but, as the level of engagement on microblogging websites is rather low, it doesn't really matter to filter everything.

Twitter shows the characteristics of a social network site but as Twitter only shows a low reciprocity between users, the primary function is more to spread information and news (Kwak, Lee et al. 2010). Going one step back, the basic function of Twitter allows the users to interact with each other. Furthermore, users can also share some personal information on a profile like part on the website. Around 80% of users update their followers on what they are doing, thinking or what happened to them. And for most of the Twitter users the number of followers they have and number of people they are following themselves is in balance. But there are some of them who show a much higher rate of followers.

On Twitter, the channels are bound to individual networks of rather weak, latent personal as well as informational ties that people choose to maintain (Haythornthwaite 2002). Maireder et all (2012) state that the way people structure their network by following people or being followed by others also determines which messages they receive or which audience they are able to communicate with. By structuring the communication using hashtags (with the # symbol placed before a key word or a full expression) or the @ sign before the name of another user, Twitterers are able to connect to a discussion stream or be very precise while communicating with someone.

An important function of Twitter is that users can forward (an action called "re-tweet") the tweets of their own followers. Re-tweeting is something that increases rapidly (Kwak, Lee et al. 2010). The re-posting of information works because members tend to follow different sets of people. This obviously also impacts the information flow via Twitter. As Kwak showed that a re-tweet can reach up to 1000 users. This also stimulates the flow as some users have a higher number of followers, but this doesn't necessarily mean that they are influential as such. While sets of people overlap and some concentrate on a particular topic, the information contained in Tweets easily moves from one set of people to the next. Moreover, an important event can stimulate a lot of tweeting (Hughes and Palen 2009), which would mean that important events are easy to detect on Twitter, as they will cause a lot of traffic. Furthermore, also offline events are closely reflected in the online world (Tumasjan, Sprenger et al. 2010).

Hashtags help users to participate very easily in a debate as the # sign transforms any word or expression into a clickable key word generating a search function for related messages. They therefore allow them to follow a specific information stream. They also help people to attract the attention of other users (Efron 2010). Furthermore, hashtags illustrate what users are talking about, what is important to them and which kind of topics they discuss: Hashtags serve "as a vehicle for otherwise unconnected participants to be able to join in a distributed conversation" (Bruns, Burgess et al. 2011, p. 49).

Even if Twitter can also be used for social purposes, there is considerable evidence that people use it significantly for information dissemination of various kinds, including personal information or statements. This personal information can be of different kind: it could be a modified re-tweet of information coming from another source, or it is a personal statement on a certain aspect, which in the same time also directs to this element, which again links personal information and information transfer. This leads to a conclusion that Twitter seems to be less a forum for political dialogue than a channel for the expression of a specific political opinion and the dissemination of news.

Public salience - Social media and public opinion

While some researchers are already turning to the “Twittersphere” as an indicator of political opinion, others have suggested that the majority of the messages are “pointless babble” (Pearanalytics 2009), others provide us with evidence that there is a link between the offline and the online world of political opinion. Therefore, the purpose of our study is to answer the question whether microblogging messages can actually inform us about public opinion and the political landscape in the off-line world. In particular, our study explores whether there is a link between online discussions and offline public opinion in the context of the 2012 Presidential election in France.

First, we examine whether Twitter is used as a vehicle for political deliberation by looking at how people use microblogging to exchange information about political issues. Second, we evaluate whether Twitter messages reflect the political preferences and the political landscape off-line in a meaningful way. Third, we are going to link data public opinion and published discourse data to see whether we can detect patterns between all three spheres.

Several previous studies have analysed online communication from a time series perspective, revealing the evolution of topics over time. Time series data of the number of blogs per day mentioning each topic were then constructed. Three common patterns for topics were found (Gruhl, Guha et al. 2004): a single spike of interest—a short period in which the topic is discussed (i.e., an increase in the number of blogs referring to it), with the topic rarely mentioned before or afterwards; fairly continuous discussion without spikes; or fairly continuous discussion with occasional spikes triggered by relevant events. It seems likely that spikes are typically caused by events in the news, but some may result from the viral spreading of jokes or information generated online.

The volume of discussion of an issue online has been used to make predictions about future behavior, confirming the connection between online and offline activities, which has been used with some success within a particular community of experts to predict stock market changes (Choudhury, Sundaram et al. 2008), influence outbreaks (Culotta 2010), or to automatically identify where and when earthquakes occur, with a high probability, from Tweets about them (Sakaki, Okazaki et al. 2010).

Twitter and political events

Political campaigns are structured by a lot of offline events such as rallies, TV debates, interviews etc. This means that the online community can reflect and talk about these events in two ways. They can

share information from news / media channels or political actors, or they just reflect on what they think about a certain event.

Work from Larson & Moe (2011) shows that Twitter is hardly used for real conversation, but to exchange and share information. The distribution of news and statements broadens the public debate as a large number of users are contributing in writing tweets. But, on the other hand, only a very small group is really active on microblogging sites. Only a smaller share of tweets actually refers to other users (Tumasjan, Sprenger et al. 2010). Especially tweets with hashtags are closely connected with offline debates and events, which somehow constitute a common reference point in the Twittersphere. Larsson & Moe (2011) as well as Bruns & Burgess (2011) also show that these offline debates are very much and often stimulated by normal mass media content.

Coming back to the earlier mentioned sentiments, it is quite interesting to see that positive tweets somehow correlate with actual figures from polls or in elections (Jürgens and Jungherr 2009). The later also shows, that external events, such as election stimulate the traffic on microblogging webpages. They are stimulated by debates, but the topics discussed on this social media forum are normally less prominently covered in mass media (Bruns, Burgess et al. 2011).

Context

The 2012 French Presidential campaign was mainly driven by the duel of the conservative candidate Nicolas Sarkozy and the socialist Francois Hollande. The French elections are organized in two rounds (unless one candidate manages to get more than 50% of votes in the first round), meaning that only the two candidates with the most votes enter the second stage.

The whole campaign period was somehow focused on two aspects, the fight between Hollande and Sarkozy on the one hand and the performance of the far right wing party Front National lead by Marie LePen. But it was relatively clear from the beginning that the two main political figures will match each other in the second round and for LePen and the other candidates it was mainly the completion for the symbolic third place. As polls showed, this race for place three was fairly open until the very last days before the first round. Eventually, LePen arrived in 3rd place. This is the reason why we decided to focus this research work on the two finalists, namely Hollande and Sarkozy.

Date	Main events
2012/02/27	Hollande announces plans for marginal tax to 75 % for incomes over one million euro per year
2012/03/02	Euro Fiscal pact signed (Sarkozy)
2012/03/06	Sarkozy TV debate with former Prime Minister Laurent Fabius
2012/03/12	Rally Sarkozy Le "grand meeting" at Villepinte
2012/03/15	Hollande TV debate with UMP party leader JF Copé
2012/03/16	Official date to hand in support list of 500 signatures for candidacy

2012/03/18	Rally Mélenchon at Bastille (Paris)
2012/03/19	Terrorist attacks in Toulouse perpetuated by Mohammed Merah
2012/03/20	Due to the Toulouse attacks the campaign stopped for one week
2012/04/04	32 proposals of Sarkozy
2012/04/05	"Elle Debate" at Science Po (Sarkozy absent)
2012/04/11	TV Show "Des Paroles et des Actes" (both candidates refused to be present in the same debate)
2012/04/12	TV Show "Des Paroles et des Actes" (both candidates refused to be present in the same debate)
2012/04/15	Rally Hollande at Vincennes / Rally Sarkozy at Concorde (Paris)
2012/04/19	Rally Hollande at Cenon
2012/04/20	Rally Sarkozy at Nice
2012/04/22	First round of election, Hollande leads (28,63 %), Sarkozy (27.18%)
2012/04/24	Sarkozy directs his campaign towards voters of Front National
2012/04/29	Rally Sarkozy Toulouse // Rally Hollande à Paris Bercy
2012/05/01	Rally Sarkozy at Trocadero // Hollande 1st of May speech
2012/05/02	TV debate between Hollande and Sarkozy
2012/05/03	Bayrou (centrist candidate) announces that he will vote for Hollande
2012/05/06	2nd round of the Presidential Election Hollande wins with 51.64% of votes / Sarkozy: 48.36%

Hypotheses

As shown above, people use the service to talk about their daily life with their friends or they seek or share for information (Java, Song et al. 2007). Furthermore, Kwak (2010) at all developed a topology on Twitter users and state that Twitter is closer to an information network than a social network. And Wu et all (2011) found that instead of acquiring directly from mass media, most people rely on an intermediate layer of opinion leaders. This finding is also supported by Larsson and Moe in their study on the Swedish elections in 2010. Twitter contributes to a broadening of the public debate, but only a small number of people constitute a substantial part of activities on the platform. The majority of users only tweets to a smaller extent and use the forum first and foremost for dissemination and not for dialogue (Larsson and Moe 2011, p. 20).

#1 Based on this, we expect to find a high correlation between media exposure of candidates and tweets directed or related to a particular candidate.

Furthermore, as Hu et all (2011) demonstrate, content is produced within the network, which directly or indirectly come from media agencies. This probably also explains the high correlation of mentions in the social media and media coverage.

#2 We therefore expect to find parallel evolutions of mass media data and social media data.

But, as we are dealing with a political event, we are aware that social media are also used as a channel for campaigning. This campaign effects should be independent from the media exposure of candidates.

#3 We expect that especially in the latter days of the campaign (closer to the election day) the social media are more a vehicle for campaigning and are not anymore related or linked to the mass media coverage of on particular candidate.

Data

During the 3 month before the 1st and 2nd round of the French presidential election we collected exhaustive data sets from the media and the social media every day. We are also using here publicly available data of representative public opinion polls collected on a daily basis.

Data measuring public opinion

The daily polls were published by Paris Match, a French magazine, on their website.² The data have been collected online on a daily basis amongst 300 to 350 people. The survey wave of the day is combined with that of the previous two days to deliver a report every night on the electoral landscape of 1,000 voters. A total of more than 30,000 people were questioned during the presidential campaign of 19 weeks covered. The rolling tracking polls or surveys smooth the effects of economic conditions, which can be very impinging in an election campaign. While a barometer of voting intentions can be compared to a series of photographs, the rolling can reproduce the film in real time of the election campaign. Taking distance from the news helps identify underlying trends, and therefore to better anticipate the scores of the respective candidates. The rolling also offers the possibility to have very large samples, and thus achieve very fine analysis by electorate.

Data showing media exposure

The measurement of media impact is based on a composite indicator that integrates editorial space (in pages or minutes) on a topic of media and audience that conveys this information to the target population 15 years and older, a tool developed by Kantar Media.³ The index of media pressure is implemented daily on a panel of 110 media (43 supports Print, 54 slices Information Radio and TV, 13 News online media) and corrected for their audience coverage.

² Source <http://www.parismatch.com/Actu-Match/Présidentielle-2012/la-présidentielle-en-temps-reel.html> (retrieved on May 7th 2012), data were collected by Ifop

³ Source <http://www.lelab2012.com/UBM.php#analyse> (retrieved on May 7th 2012), data were collected from KantarMedia

Data from the social media domain

The main focus for the French Presidential Elections was the analysis of the content of Twitter as it is the social network that channels the biggest volumes of posts in real-time⁴.

When multiple repetitive acronyms appear in our research, we apply a polysemic filter to our search engine to make sure that these acronyms are only taken into account when we are certain that they related to our theme.

However, this filtering is only possible when words are associated with these abbreviations. Therefore, we have also developed a corpus of words that are most often related to our theme in order to reject all unrelated messages using these abbreviations. Finally, we only count once the messages that are perfectly identical and published several times by the same (this initiative was implemented to remove the phenomena of "spams"). Once we have collected all the messages and applied polysemic filtering and "de-spamming" to our research, we then analyse the sentiment contained in every messages.

Sentiment analysis

Sentiment analysis is useful for research into online communication because it gives researchers the ability to automatically measure emotion in online texts. The research field of sentiment analysis has developed algorithms to automatically detect sentiment in text (Pang and Lee 2008). There are many IT tools on the market that claim to measure the tone or feeling of messages exchanged in social media channels. Most of the available tools that we have tested are either primarily designed for commercial reputation/image issues (products or brands) and have proven to be disappointing when it comes to sentiment analysis.

While some identify the objects discussed and the polarity (positive, negative, or neutral) of sentiment expressed about them (Gamon, Aue et al. 2005), other algorithms assign an overall polarity to a text, such as a movie review (Pang, Lee et al. 2002). Three common sentiment analysis approaches are full-text machine learning, lexicon-based methods, and linguistic analysis (Pang, Lee et al. 2002; Witten and Frank 2005; Pak and Paroubek 2010).

This rather "unreliable" reality compelled us to develop, in a partnership with Vigiglobe sprl, a new customised tool that is perfectly dedicated to sentiment analysis of messages based on political and social themes.

This analysis tool is based on text mining and "machine learning" algorithms that classifies the messages according to three tones:

- Positive: ie. posts in which the author acknowledges a real quality to the subject or approves of their action.
- Informative: ie. posts in which the author do not share his opinions or any views at all. He/she only relays information about the subject researched.

⁴ In France, in November 2011, 3% of people had a twitter account and around 67 % used this account at least once a week Data collected by TNS in November 2011. Considering the rapid growth of Twitter penetration in Europe, it is fairly likely that this percentage has increased since then.

- Negative: ie. posts that express a critique or negative view towards, the subject or their action. The author explicitly shares his disapproval or irony with regards to the subject.

These complex algorithms study the statistical relationships between words or expressions and are based on collections/corpus of sentences, sometimes whole paragraphs (and not adjectives' dictionaries), in order to decipher, understand and interpret the tone of a particular message. To take into account the undeniable fact that people express themselves differently in context and time, we have developed a thematically corpus for the French Presidential Elections) by using native speaking teams experienced in the field of open questions codification. This corpus is continuously evaluated for accuracy. This step allows also the continuous improvement of the tool as internet users have a great tendency to create/invent words and acronyms to express a sentiment on a specific topic.

The evaluation process uses both human and machine scoring of the same content. It is based on accuracy scores, that is the number of correctly predicted content divided by the size of the whole dataset. Human scoring is performed in a single pass by people who did not participate in the training process (i.e in building the corpus). This human classification is to some extent different from an individual to another due to the inherent subjectivity of perception. The results of the human classification are then compared with the machine scoring to determine accuracy. On the French classification of political content, this score reached 75% of accurate sentiment analysis.

Results

We begin by examining general features of our data by looking at daily Twitter activity and media exposure of candidates. In a second step, we are going to see, whether there is a qualitative link between topics discussed in the social media environment and external events. In a third step we will link the social media data with the results of the daily opinion polls data. In order to analyse the data from the different sources we standardized all data sets using the z-transformation. The advantage of the z-transformation is that the structure of the data remains untouched, which enables us to compare the evolution.

Relationship between Twitter data and Media

The graph below shows that the data for the total amount of Tweets and the media exposure for candidate Hollande (a) and candidate Sarkozy (b) from February 1st to May 6th 2012. The first round took place on April 22 and the 2nd round on March 6.



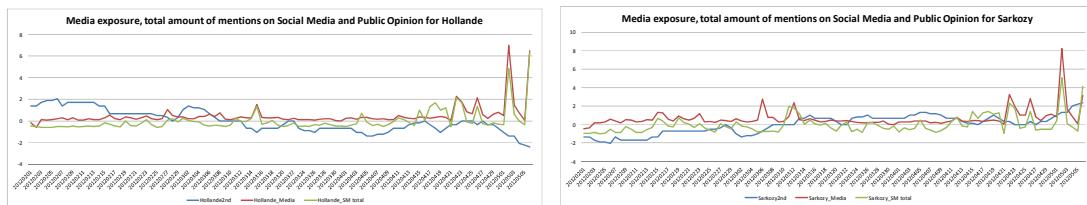
The graphs indicate two major aspects, the Twitter based data more or less follow the media data, or in other words the amount of tweets is connect with the media coverage for the candidates. The correlation for the raw data correlation is 0.2 for Sarkozy and 0.3 for the data of Hollande.

		Overall	February	March	April	May
		Average	Average	Average	Average	Average
Public opinion data	Sarkozy	45.0	43.4	45.3	45.9	47.5
	Hollande2nd	55.0	56.4	54.7	54.2	52.5
Media coverage	Sarkozy_M	587	427	458	574	2064
	Hollande_M	494	255	293	496	2641
Social media	Sarkozy_P	130760	82659	120180	151736	313036
	Hollande_P	86166	31969	56615	117940	341933
Sentiment	Sarkozy_positive	12623	10902	36211	5593	14075
	Sarkozy_negative	29926	27630	72043	14491	26281
	Sarkozy_positive-negative ratio	0.39	0.38	0.54	0.11	0.09
	Hollande_positive	9717	5432	40384	2876	14534
	Hollande_negative	21154	14618	79377	8473	26141
	Holland_positive-negative ratio	0.43	0.38	0.58	0.09	0.10

The second finding relates to the week before the first round of the election (April 15 to April 22). The media coverage remains low, but for both candidates the social media data increase significantly.⁵

Social media, media and polls

In order to show the relationship between media coverage – traditional and web 2.0 – we take a look at the graph including all three variables.



⁵ Looking into the word cloud related to this period, one can see that the traffic seems to be related to campaigning of the parties, as for all candidates (not only Hollande and Sarkozy) users tweet invocations to vote.

Based on this graph, the results of polls and both media sources are not as such related with each other. Whereas the mentions in both media sources show a similar evolution for both candidates, the poll results for Hollande show a negative evolution, whereas Sarkozys' numbers go up.

Modeling social media and public opinion polls

As voting intentions and answering on opinion polls are strongly related to political engagement as well as certain political attitudes, we also need to model this dimension into the media data.

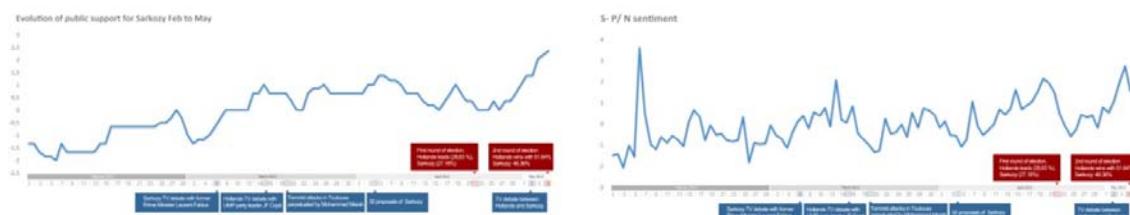
The data from Twitter were also coded (as described above) with respect to positive and negative sentiments as well as neutral messages. We treat the tweets as the aggregate political opinion of the users of this social media platform.

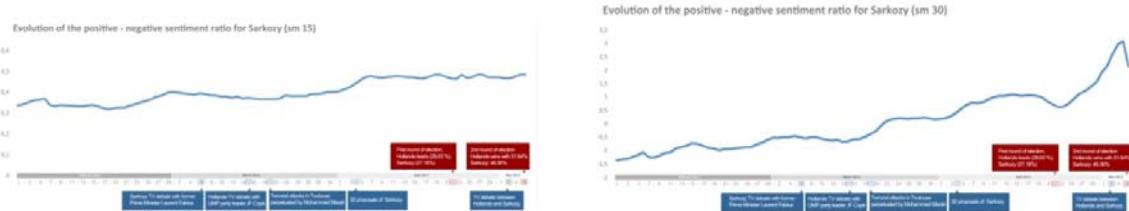
In order to calculate an aggregated public sentiment we calculate a ratio score for positive and negative messages (O'Connor, Balasubramanyan et al. 2010). We derive day to day sentiment scores by counting positive and negative tweets, which have been coded as described above. We defined the sentiment score x_t on day t as the ratio of positive versus negative messages for each candidate and obtain the relative frequency for the sentiment.

		Overall	February	March	April	May
Sarkozy		Average	Average	Average	Average	Average
Media coverage	raw	587	427	458	574	2064
	smooth for 7 days	630	430	1794	133	564
	smoothed for 30 days	769	379	1794	51	327
Sentiment ratio	Positive/ Negative Ratio	0.39	0.38	0.54	0.11	0.09
	Positive/ Negative Ratio 7 day	0.40	0.38	0.58	0.03	0.06
	Positive/ Negative Ratio 30 day	0.42	0.40	0.58	0.01	0.02
Hollande						
Media coverage	raw	494	255	293	496	2641
	smooth for 7 days	600	284	3295	88	580
	smoothed for 30 days	813	295	3295	38	455
Sentiment ratio	Positive/ Negative Ratio	0.43	0.38	0.58	0.09	0.10
	Positive/ Negative Ratio 7 day	0.44	0.39	0.58	0.04	0.07
	Positive/ Negative Ratio 30 day	0.45	0.42	0.59	0.01	0.03

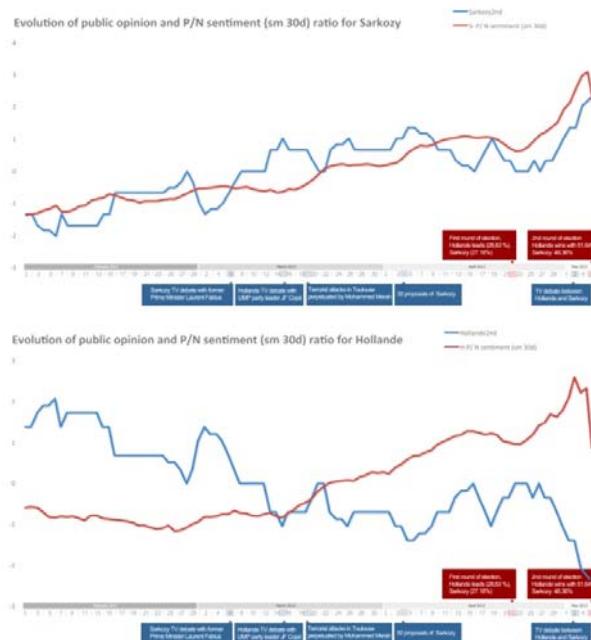
As the day to day sentiment ratio is relatively volatile, compared to traditional polls we smooth the sentiment ratio using a moving average over a window of 7 days and 30 days. Smoothing is of course a critical issue as it causes a much slower response to recent changes. Instead it shows a more consistent behaviour over a longer period of time. Too much smoothing could possibly cover changes in the sentiment analysis.

Focusing on the sentiment ratio, and correcting it for the noise in the data (using a moving average approach) we are able to show that the curves for positivity are strongly linked to the poll results.

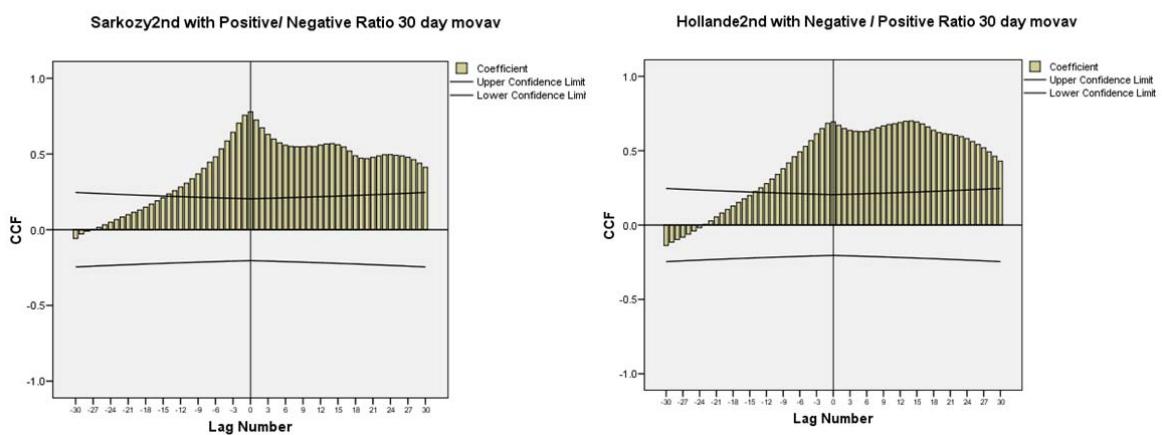




Below you can find the smooth data which shows both, the social media data and the poll results. Whereas we see a parallel trend for Sarkozy, the data for Hollande reveal a different trend.

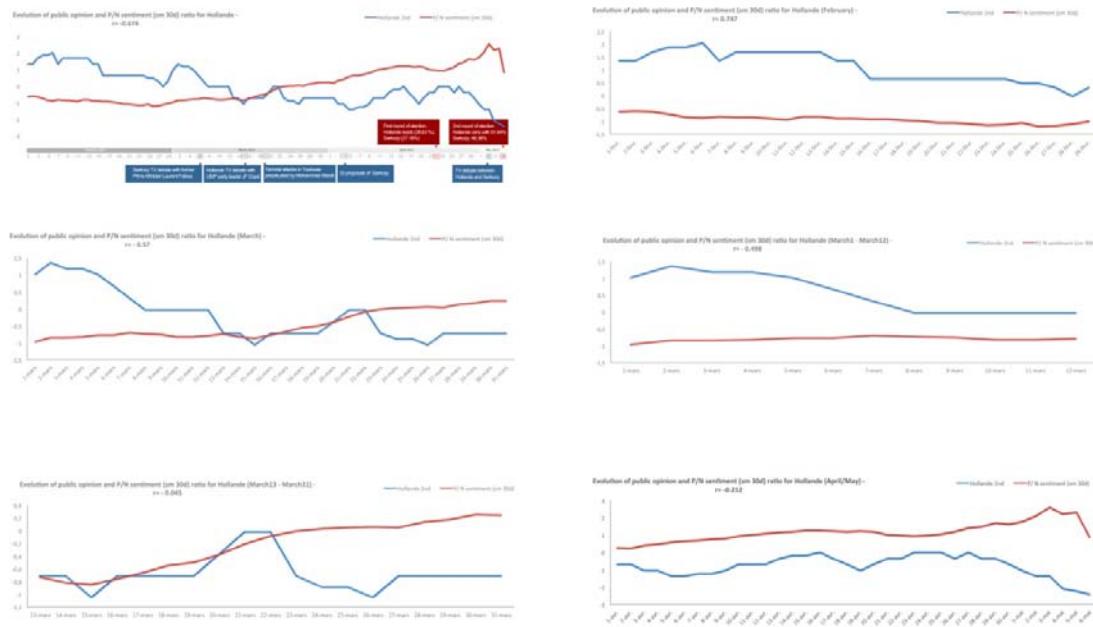


While looking into the data, we also checked the cross correlation for the two concepts. While using a cross-correlation factoring we can also assume the direction of the relationship. The sensitivity of the sentiment-poll correlation using the smoothed data (30 days) is shown below. Cross-correlation is higher for tweets, meaning that the tweets leading the poll.



The graph above shows the lagged cross-correlation. On the left side of the Lag=0 you can see that the correlation between social media and polls is rapidly decreasing. On the right side of Lag=0, the correlations are decreasing as well, but much slower. This means that the correlation is lagged in one direction, from text messages to poll results. Or in other words, the polls follow the tweets.

But as already indicated above, the results for Hollande are not in line with our hypothesis. But, maybe it helps to take a closer look into the data. In the week between March 5 and March 12 something changes the discourse structure.



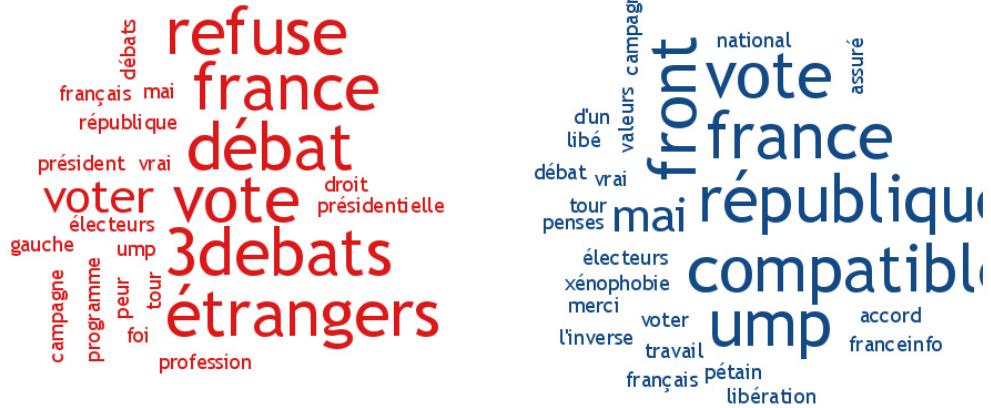
We know from above that after Sarkozy appeared on television (March 6 and March 12) the support for Sarkozy increased significantly. On the other hand, the trend for Hollande remained negative. The overall correlation between poll results and sentiment ratio is negative and shows a rather high correlation. Splitting the timeline in different sections, one can find a slightly different result, as the trend until the TV appearance of Sarkozy we find a strong positive correlation. The effect for month is again negative and rather strong, whereas the correlation in the April until the Election Day on May 6 is weak and negative.

These findings for the second week of March are somehow supported by the data on media exposure for the two candidates. In this period the mentions of both candidates in the media remains rather stable, meaning only the social media domain shows a totally different evolution for Hollande.

Hollande	Sarkozy
Mai 1 st	



The final debate



Again using the richness of the data set, we can see in the qualitative information (word clouds and phrases) that especially shortly before the Election days (April 22, May 6) new #hashtags appear. We are tempted to conclude from this, that social media become a campaign channel. Most of these new hashtags are by its nature neutral, but transport a clear sentiment, which works against one of the two candidates. The supporters of Hollande are much more active in this respect.

Conclusion

In the paper we presented the results of a time based analysis for the French Presidential Election 2012. The data reveal that data and sentiments on social media platforms can be used to get a better understanding of public opinion.

Our analysis suggests that mass media is still at the core of reporting also in the social media sphere. But this exercise also shows that the amount of mentions within the social media is much higher and quicker than traditional media as it is organized as a many to many communication face that is able to spread the news or discussing streams effectively.

The amount of mentions for candidates is highly correlated with their appearance in classical media streams and social media function as new agora for information exchange. Of course, one has to be

quite carefully with directing the effect from one to the other source of information. It could also be the case that the public salience medium (social media) stimulates the media salience. But, as the analysis indicates, most of the buzz happens in parallel, and in really short time, so that it is more likely that media content has been used to create noise in the social media sphere.

On the other hand, we could show that the closer the Election Day the more likely the relationship between media and social media collapse and social media develop its own dynamic, in amount and tone. We understand this as a clear hint that social media nowadays are more and more an instrument to campaign and mobilize voters.

We also saw, that it is not the amount of mentions on social media, but their qualitative nature, which helps us to establish a link between public salience on social media and public opinion polls. While the results do not come without caution, it is encouraging to see that a careful text mining analysis of online social media content can give us valuable information of public opinion trends which evolves as the campaign moves forward. This type of sentiment analysis needs to be further developed and continuously adapted and updated as citizens invent new words and concepts as the campaign goes by.

Furthermore, as it is not the total amount of tweets that helps us to establish a link between public opinion measures and discourse on social media channels, one needs to also think about the enormous amount of qualitative. The daily word clouds, the profile and the activities of different users in combination with sentiment will provide us with even more information on, how users contribute to the formation of public opinion. The creation of new #hashtags is only one interesting finding, the dominance of specific phrases throughout the whole campaign period is another (terms such as PS, La France Forte etc). On the other hand, the uniformity of terms used to communicate on Twitter also shows that the participation in this kind of communication is less interactive and not focused on generating any response.

On the other hand, low engagement practices such as re-tweeting have a central role in information propagation process allowing messages to reach a larger number of users and therefore bringing news from something known within specific circles to something widely known within the society at a large (Rossi, Magnani et al. 2011). While traditionally media events left the audience alone, when it comes to contemporary media events supported by social media, the audience is made visible and it plays an active role within the propagation process. These findings confirm that the focus on social media content offers quite interesting insights in how discussions in our societies emerge and are framed.

Limitations

The data have been gathered from different sources and are not based on one single and interlinked research design. In our paper, we used the online public opinion poll as the reference category and did not address the mode effect here.

Another point that needs to be mentioned here is the fact that we did not look at topics / issues communicated by the candidate and how this is reflected in the social media or media coverage. This

could add an important dimension, because it will clarify the so far blurred picture of how the salience of public opinion is related to an important measure such as a choice of vote.

Furthermore, the analysis could be enriched by a specific analysis of users' behaviours (and perhaps social media strategies) during this very long campaign. We indeed did not focus yet on the interaction between users, the position of single users within a network and their potential online influence. Due to time restrictions, we did not pay attention to the whole Re-tweet aspect, which could help us to gain deeper insights in the mechanisms of how discourses on social media are linked to media coverage. It could be interesting to evaluate the interrelation of sentiments of social media and classical media in order to link public salience and media salience.

Last but not least, we need to think about using more sophisticated methods of time series modeling in order to really show the linkage between media, social media and public opinion polling.

Acknowledgments

We need to thank Vigiglobe for preparing the social media data base and Kantar Media for providing us with the media data. Special thanks to Caroline Ashdown for mediating between the different institutes and the authors of this paper.

Literature

- Boyd, D. M. (2004). Friendster and publicly articulated social networks. Proceedings of ACM Conference on Human Factors in Computing Systems, Vienna, April 24-29.
- Boyd, D. M. (2007). The significance of social media. Blog talks reloaded: Social software & cases. T. N. Burg and J. Schmidt. Norderstedt, Books on Demand: 15 - 30.
- Boyd, D. M. and N. B. Ellison (2007). "Social network sites: Definition, history, and scholarship." Journal of Computermediated Communication 13(1).
- Bruns, A. (2011). "How long is a Tweet? Mapping Dynamic Conversation Networks on Twitter using Gawk and Gephi." Information, Communication & Society: 37 - 41.
- Bruns, A., J. Burgess, et al. (2011). "Mapping the Australian Networked Public Sphere." Social Science Computer Review 29(3): 277 - 287.
- Brustein, J. (2010). Nation's political pulse, taken using net chatter. New York Times. New York.
- Choudhury, M. D., H. Sundaram, et al. (2008). Can blog communication dynamics be correlated with stock market activity? Proceedings of the 19th ACM Conference on Hypertext and Hypermedia New York: ACM Press.
- Coeckelbergh, M. (2010). WHAT ARE WE DOING? MICROBLOGGING, THE ORDINARY PRIVATE, AND THE PRIMACY OF THE PRESENT. Enschede, Emerald Group Publishing Limited.
- Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages.
- Delli Carpini, M. X., F. L. Cook, et al. (2004). "Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature." Annual Review of Political Science 7: 315 - 344.
- Delli Carpini, M. X. and S. Keeter (2003). The Internet and an informed citizenry. The civic web: Online politics and democratic values. D. M. Anderson and M. Cornfield. Lanham, MD, Rowman & Littlefield: 129 - 153.
- Efron, M. (2010). Hashtag retrieval in a microblogging environment. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press.
- Gamon, M., A. Aue, et al. (2005). "Pulse: Mining customer opinions from free text." Lecture Notes in Computer Science 3646: 121 - 132.
- Gruhl, D., R. Guha, et al. (2004). Information diffusion through Blogspace. Proceedings of the 13th International Conference on World Wide Web.
- Hawn, C. (2009). "Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media are Reshaping Health Care." Health Affairs 28(2): 361 - 368.
- Haythornthwaite, C. (2002). "Strong, Weak and Latent Ties and the Impact of New Media." The Information Society 18(5): 385 - 401.
- Hu, M., S. Liu, et al. (2011). Breaking News on Twitter.
- Hughes, A. L. and L. Palen (2009). "Twitter adoption and use in mass convergence and emergency events." International Journal of Emergency Management 6(3/4): 248 - 260.
- Jansen, H. J. and R. Koop (2005). "Pundits, ideologues, and ranters: The British Columbia election online." Canadian Journal of Communication 30: 613 - 632.

- Java, A., X. Song, et al. (2007). Why we twitter: understanding microblogging usage and communities. Proceedings WebKDD/SNA-KDD'07, ACM.
- Jungherr, A., P. Juergens, et al. (2011). "Why the Pirate Party Won the German Election in 2009 or the Trouble With Predictions: A Response to Tumasjan, Sprenger, Sandner and Welpe "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment". " Social Science Computer Review, published 25 April 2011.
- Jürgens, P. and A. Jungherr (2009). Twittering Dissent: Social Web Data Streams as Basis for Agent Based Models of Opinion Dynamics. 11th General Online Research Conference, GOR 09: Proceedings, Wien, p. 81.
- Kushin, J. M. and M. Yamamoto (2010). "Did Social Media Really Matter? College Students' Use of Online Media and Political Decision Makingt in the 2008 Elections." Mass Communication and Society **13**: 608 - 630.
- Kwak, H., C. Lee, et al. (2010). What is Twitter, a Social Network or a News Media? WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
- Lampe, C., N. Ellison, et al. (2006). A Face(book) in the crowd: Social searching vs. social browsing. Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (pp. 167-170). New York: ACM Press.
- Larsson, A. and H. Moe (2011). "Studying political microblogging: Twitter users in the 2010 Swedish election campaign." New Media & Society.
- Lupia, A. and T. S. Philpot (2005). "Views from inside the Net: How websites affect young adults' political interest." The Journal of Politics **67**: 1122 - 1142.
- Maireder, A., J. Ausserhofer, et al. (2012). Mapping the Austrian Twittersphere. How politicians, journalists and political strategists (inter-)act on Twitter. CeDEM12. Conference for E-Democracy and Open Government. P. Parycek and N. Edelmann. Krems, Danube University Krems, 3-4 May 2012: 151 - 163.
- McLuhan, M. (1964). Understanding Media: The Extension of Man. New York, McGraw-Hill Book Company.
- Naaman, M., H. Becker, et al. (2011). "Hip and Trendy: Characterizing Emerging Trends on Twitter." JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY **62**(5): 902 - 918.
- Naaman, M., J. Boase, et al. (2010). Is it really about me? Message content in social awareness streams. Proceedings of the 2010 ACM conference on Computer supported cooperative work. Savannah:ACM.
- O'Connor, B., R. Balasubramanyany, et al. (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.
- Pak, A. and P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC 2010 Paris: European Language Resource Association.
- Pang, B. and L. Lee (2008). "Opinion mining and sentiment analysis. Foundations and Trends." Information Retrieval **2**: 1 - 135.
- Pang, B., L. Lee, et al. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Morristown, NJ: Association for Computational Linguistics.

- Papacharissi, Z. and A. Mendelson (2008). Friends, networks and zombies: The social utility of Facebook. Annual meeting of the Association of Internet Researchers, Copenhagen, Denmark.
- Pearanalytics (2009). Twitter study.
- Putnam, R. D. (2000). Bowling alone: The collapse and revival of American community. New York, Simon & Schuster.
- Rossi, L., M. Magnani, et al. (2011). #rescatemineros: global media events in the microblogging age. IR 12.0. Association of Internet Researchers, 10-13 October, Seattle, US.
- Royal, C. (2008). User-generated content: How social networking translates to social capital. Annual meeting of the Association for Education in Journalism and Mass Communication, Chicago, IL.
- Sakaki, T., M. Okazaki, et al. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. Proceedings of the 19th International Conference onWorldWideWeb (WWW'10). NewYork: ACM Press.
- Severin, W. J. and J. W. Tankard (2001). Communication Theories: Origins, Methods, and Uses in the Mass Media. New York, Addison Wesley Longman.
- Tumasjan, A., T. O. Sprenger, et al. (2010). "Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape." Social Science Computer Review 12 December 2010.
- Tweetminster (2010). "Is word-of-mouth correlated to General Election results? The results are in. Retrieved March 25, 2012."
- Vargo, C. (2011). National television news and newspaper as media salience, twitter as public salience: an agenda setting effect analysis, The University of Alabama, Tuscaloosa Alabama.
- Verbeek, P. (2005). What Things Do: Philosophical Reflections on Technology, Agency and Design. Pennsylvania, Pennsylvania State University Press.
- Williams, C. and G. Gulati (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. Annual meeting of the American Political Science Association (pp. 1–17). Boston, MA: APSA.
- Witten, I. H. and E. D. Frank (2005). Data mining: Practical machine learning tools and techniques. San Francisco, Morgan Kaufmann.
- Wu, S., P. Hofman, et al. (2011). Who says what on twitter? Proceedings WWW'11, ACM 2011.