Politics and Big Data

Nowcasting and Forecasting Elections with Social Media

Andrea Ceron, Luigi Curini and Stefano Maria Iacus

ROUTLEDGE

Politics and Big Data

The importance of social media as a way to monitor an electoral campaign is well established. Day-by-day, hour-by-hour evaluation of the evolution of online ideas and opinion allows observers and scholars to monitor trends and momentum in public opinion well before traditional polls. However, there are difficulties in recording and analyzing often brief, unverified comments while the unequal age, gender, social and racial representation among social media users can produce inaccurate forecasts of final polls. Reviewing the different techniques employed using social media to nowcast and forecast elections, this book assesses its achievements and limitations while presenting a new technique of "sentiment analysis" to improve upon them. The authors carry out a meta-analysis of the existing literature to show the conditions under which social media-based electoral forecasts prove most accurate while new case studies from France, the United States and Italy demonstrate how much more accurate "sentiment analysis" can prove.

Andrea Ceron is Assistant Professor of Political Science at Università degli Studi di Milano, Italy.

Luigi Curini is Associate Professor of Political Science at Università degli studi di Milano, Italy.

Stefano Maria Iacus is Full Professor of Mathematical Statistics and Probability at Università degli Studi di Milano, Italy.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Politics and Big Data

Nowcasting and Forecasting Elections with Social Media

Andrea Ceron, Luigi Curini and Stefano M. Iacus

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2017
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Names: Ceron, Andrea, author. | Curini, Luigi, author. | Iacus, Stefano M. (Stefano Maria), author.
Title: Politics and big data : nowcasting and forecasting elections with social media / Andrea Ceron, Luigi Curini and Stefano M. Iacus.
Description: Abingdon, Oxon ; New York, NY : Routledge, [2017] | Includes bibliographical references.
Identifiers: LCCN 2016034245| ISBN 9781472466662 (hbk) | ISBN 9781315582733 (ebk)
Subjects: LCSH: Election forecasting. | Social media—Political aspects. | Big data—Political aspects. | Internet in political campaigns.
Classification: LCC JF1048 .C47 2017 | DDC 324.900285/57—dc23
LC record available at https://lccn.loc.gov/2016034245

ISBN: 978-1-4724-6666-2 (hbk)
ISBN: 978-1-315-58273-3 (ebk)

Typeset in Times New Roman
by Apex CoVantage, LLC

Contents

List of figures vi
List of tables viii

Introduction 1

1 Social media electoral forecasts: an overview 5

2 From noise to signal in sentiment and opinion analysis 38

3 Nowcasting and forecasting electoral campaigns: evidence from France, the United States and Italy 68

4 Leaders, promises and negative campaigning: digging into an electoral campaign through social media 105

5 Social media and electoral forecasts: sources of bias and meta-analysis 132

6 Conclusion: "To predict or not to predict?" Future avenues of social media research within and beyond electoral forecasts 154

Postscript 171
Index 175

Figures

2.1 Needs and solutions in textual analysis 45
2.2 The data generating process in topic models 49
2.3 Intuitive reason why, when the noise D_0 category is dominant in the data, the estimation of $P(S \mid D)$ is reasonably more accurate than the estimation of counterpart $P(D \mid S)$ 52
2.4 The estimated sentiment toward Donald Trump 62
2.5 Reasons for positive sentiment toward Donald Trump 63
2.6 Reasons for negative sentiment toward Donald Trump 64
3.1 First round of the 2012 French presidential election: vote share predictions according to electoral surveys and SASA by presidential candidates (SASA) 72
3.2 Flow of preferences expressed on Twitter during the electoral campaign for the second round of the 2012 French presidential election (SASA) 74
3.3 Second round of the 2012 French presidential election: vote share predictions according to electoral surveys and SASA by presidential candidates 75
3.4 2012 US presidential election: daily social voting intentions according to SASA 77
3.5 Estimates of the lead-lag analysis in the short term 87
3.6 General expectations on Twitter about the outcome of the 2012 centre-left primary election (SASA) 89
3.7 Centre-left primary election, first round: candidates' share of votes according to SASA and comparison with actual results 90
3.8 First round of the 2012 Italian centre-left primary elections: absolute difference between final results and predictions according to survey polls and SASA 91
3.9 Italian Democratic Party primary election: daily social voting intentions according to SASA 96
3.10 Italian Democratic Party primary election: daily evolution of the level of negative campaigning in Twitter discussion over candidates 98

4.1 Flow of preferences expressed on Twitter (SASA) during the last month of electoral campaign for the five main coalitions in the 2013 Italian general elections 109
4.2 Predicted vote share for the centre-right and centre-left coalition according to survey polls throughout the 2013 Italian electoral campaign 112
4.3 Comparison between actual vote share and, respectively, predictions according to survey polls, instant polls and SASA 113
4.4 Evolution of the different strategies across the campaign 117
4.5 Marginal effect of negative campaign toward rivals conditional on the magnitude of negative campaign suffered (with 95% confidence interval) 120
4.6 Marginal effect of distributive/clientelistic policy promises conditional on the party making the promise (with 95% confidence interval) 121
4.7 Evolution of sentiment toward Italian political leaders before the 2013 Italian general elections (SASA) 123
4.8 Evolution of sentiment toward Italian political leaders after the 2013 Italian general elections (SASA) 124
5.1 Evolution over time of the MAE of social media leaders' approval (SASA) 135
5.2 The evolution of the MAE over time as the election date approached (SASA) 138
5.3 Predicted and actual vote shares related to the first round of the 2012 French legislative elections 140
5.4 Marginal effect of the number of tweets on the MAE as turnout changes (with 90% confidence interval) 142
5.5 Marginal effect of number of posts on the MAE at different levels of turnout (with 95% confidence interval) 150
6.1 Distribution of ideological self-placement of citizens by frequency of SNS usage 159
6.2 Distribution of ideological self-placement of citizens by perceived importance of SNS for expressing own views 159
PS.1 2016 US presidential election: daily social voting intentions according to iSA 173

Tables

1.1 A typology of methods employed to perform social media-based electoral forecast 16

2.1 What characterizes text analysis techniques 42

2.2 Example of document-term matrix 43

2.3 Comparison of individual and aggregated classifiers 58

3.1 Temporal evolution of the main topics of the US presidential campaign according to social media conversations 78

3.2 Reaction to the presidential debates during the US electoral campaign according to survey polls and Twitter (SASA) 79

3.3 2012 US presidential election: accuracy of the predictions. 83

3.4 Comparison of the accuracy of Twitter forecast made through mentions, automated sentiment analysis and SASA method (2012 US presidential election, popular vote) 84

3.5 Estimates of the lead-lag analysis in the short term 86

3.6 Comparison of the accuracy of Twitter forecast and surveys polls in the Italian primary election of the centre-left coalition (first round) 92

3.7 Comparison of the accuracy of Twitter forecast and surveys polls in the Italian primary election of the centre-left coalition (second round) 93

3.8 Comparison of the accuracy of Twitter forecast made through mentions, automated sentiment analysis and SASA method (Italian primary election of the centre-left coalition, first round) 93

3.9 Italian Democratic Party primary election: contrasting the actual votes of the candidates with Twitter (SASA) predictions (both weighted and non-weighted) 97

4.1 Effect of different campaign strategies on a party’s share of voting intentions 119

4.2 Explaining surveys voting intentions according to leaders’ positive sentiment 126

5.1 Average difference and correlation of leaders’ popularity ratings between mass surveys and social media 134

5.2 Fractional logit of the MAE 141

5.3 Number of predictions by election type 144
5.4 List of countries and number of predictions 145
5.5 Determinants of the accuracy of social media predictions: fractional logit of the electoral forecast MAE 148
6.1 Weighting social media predictions: fractional logit of the electoral forecast MAE 157
6.2 Distribution of ideological self-placement of Italian voters versus subsample active on social media 158
6.3 Survey data as determinants of the accuracy of social media predictions: fractional logit of the electoral forecast MAE 163

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction

> The best technology is human-empowered and computer-assisted.
>
> (Gary King, May 18, 2016)

Political and social sciences are undergoing a complex process of revolution, which has been summarized under the label of “Big Data revolution” and, indeed, Big Data certainly play a crucial role in such process.

Big Data are those labeled, for strange reasons, with the capitalized “Big”. Nevertheless, they are still “Data”, and this word too deserves a capital letter to suggest that good statistical techniques are required in order to extract meaningful results from such sources. Moreover, Big Data does not even mean *just* a data collection with a very large-N (Dalton 2016). For example, a very large survey of citizen participation cross-nationally or aggregating several surveys is not, strictly speaking, Big Data. Rather, Big Data typically involves the collection of indirect evidence of massive size from diverse and often unconventional sources. In this respect, the main features of Big Data can be summarized according to three main aspects (Couper 2013; Daas et al. 2012): 1) “*volume*” – they exceed the capacity of traditional computing methods to store and process them; 2) “*frequency*” – data come from streams or complex event processing, size per unit of time matters; and 3) “*unpredictability*” – data come in many different forms: raw, messy, unstructured, not ready for processing, not relational and so on. Therefore, they require structure before use by conventional analysis methods or might call for the application of new analytic methods of data processing.

Big Data are also sometimes called “organic data” (Groves 2011) to express the idea that these data are generated, most of the time, automatically by systems, computers or probes or humans interacting through digital devices. In fact, the main types of these data can be summarized in terms of their source as follows:

- *Administrative data*: stored in databases generated by persons or organizations for regulatory or other government activities. In most cases these data are confidential, but in the last years, with the advent of the “Open Data” philosophy, more and more data are becoming available for public use although none of these data are usually collected or designed for scientific purposes.

- *Transaction data*: they are generated through high frequency financial transactions on the stock markets, e-commerce transactions or loyalty card–type transactions at shops, phone records, Web surfing and so forth; these data, originally collected to increase the user experience, are clearly used for secondary scopes like marketing and profiling in general. While these data may have a variety of different structures, the nature itself is well defined because they are collected for specific purposes. Their main issues consist in how to narrate them.
- *Social media or social networking data*: these data are created by explicit actions of people and by interactions with others. The purpose of these data is different and depends on the user and the social media, unpredictable in terms of time, space, volume and content, and highly unstructured (for the same user they can be "likes", pokes, winks, photos, quotes, etc.), identity and uniqueness of the user is hard to assess. Nevertheless, these data convey a lot of information to social scientists.

The broadening of Internet penetration and the growing number of worldwide citizens active on social networking sites like Facebook and Twitter pushed the "Big Data revolution" further. In this new "digital world", citizens share information and opinions online, thereby generating a large amount of data about their tastes and attitudes. This information can then be successfully exploited to study more in depth the formation and evolution of public opinion (Schober et al. 2016) – also by integrating online these data with more traditional ones (Couper 2013).

In this book, we therefore focus on the third type of Big Data, that is social media data. These data are, by far, the most unstructured. They are volatile (people appear and disappear, and this also applies to the topics debated online), but at the same time they are fast growing and high frequency data. In addition, these data are mainly composed of digital texts that are the expression of natural and informal language and contain a lot of noise.

In view of that, we must keep in mind that the Big Data revolution is related not only to data sources. The evolution of our societies toward a digital world is a necessary premise. However, the methodological contribution of information technology, which allows us to gather and store huge quantities of data, processing them at an incredibly fast rate, and the new developments in statistics and political methodology, particularly in the field of text analysis, are also important in performing such revolution. In this regard, we devote one part of the book to discuss the methodological requirements of Big Data analysis and to show the recent improvements that make the analysis of social media possible and reliable. In particular, we present a modern technique of sentiment analysis tailored to the analysis of social media comments (Ceron et al. 2016), and we demonstrate how such technique allows us to extract accurate information from these rich sources of data. The method we propose relies neither on statistical methods nor human coding alone. The intuition is that we can combine the superior ability of humans to read and understand the meaning of natural language with the superior ability of computers and algorithms to aggregate data into reliable categories.

Throughout the book, this new technique of sentiment analysis is used to analyze the opinions expressed during several electoral campaigns held in countries such as the United States, France and Italy, for the purpose of nowcasting the evolution of the campaign and to produce electoral forecasts. In the field of social media forecast, it is quite rare to have a benchmark to assess the validity of the analysis. The case of electoral competition, however, is one of the few exceptions in this regard. As we discuss at length, this is an extremely rare case in which it is possible to have a precise benchmark to compare the research output with the actual votes count.

The six chapters of the book provide a comprehensive analysis of the phenomenon of social media-based electoral forecasts, from both a methodological and an empirical point of view. In this regard, Chapter 1 presents a detailed and reasoned review of the literature on social media, focusing on the *seven main streams* of research that have emerged over time. In detail, it discusses studies on the impact of social media on political engagement as well as the relationship with collective action and public policy. Analogously, the relationship between social media and old media or the role of social media in e-campaigning (useful also to estimate the position of political actors) is addressed. Particular attention is devoted to studies on the emotions and on the opinions expressed on social media as well as on the usage of online data to produce a wide variety of economic, social and political forecasts. Finally, this chapter starts to categorize the existing studies on social media and electoral forecast, describing the main features of different approaches but also highlighting the limitations of this stream of research.

Chapter 2 is more methodological. Starting from a brief introduction to several techniques of text analysis, this chapter describes iSA (*integrated* sentiment analysis), which is the supervised aggregated sentiment analysis (SASA) technique that is used to produce our electoral forecasts and that allows us to distinguish the hard signal from the large portion of noise that affects social media conversations. Chapter 2 also describes the free iSAX package for the R statistical environment. Following the guidelines detailed in Chapter 2, this package can be used (for noncommercial purposes) to dig into the opinions of social media users and to replicate one of the analyses discussed in the book about the sentiment toward Donald Trump.

Chapter 3 is the first empirical chapter. It focuses on elections held in France, the United States and Italy, providing evidence of how the electoral campaign can be profitably nowcasted and the final results of the election can be successfully predicted. Per each of the eight analyses, we fully report details on the data sources, the period of analysis, the data gathering process (including the full list of keywords), the number of comments analyzed and the sentiment analysis technique adopted. We also report the date of the election and that of the last publicly released forecast along with a URL or a reference for such forecast.

Focusing on the Italian case, and by analyzing the popularity of Italian leaders between 2012 and 2013 as well as the online voting intentions expressed during the campaign for the 2013 Italian general election, Chapter 4 looks at the other side of the coin. This chapter, in fact, shows how the information available on social media can also be used to analyze (and nowcast) the effectiveness of campaign strategies and to provide new insights on the study of traditional political science topics such as the role of political leaders and their valence endowment, the impact of policy promises and that of negative campaigning.

Chapter 5 points more directly to the electoral forecasts and reports two statistical analyses to investigate what elements increase the accuracy of social media-based predictions by reducing the mean absolute error. First, with respect to the French legislative election, we compare the actual results in 46 local districts with our social media predictions, showing that the accuracy is higher when there is more consistency between online declarations and actual behavior. In other words, having more information on citizens' voting preferences decreases the error only when the turnout rate is sufficiently high whereas having more information increases the error when voters tend to abstain at a higher rate. Second, we perform a meta-analysis of 239 electoral forecasts related to 94 different elections, held between 2007 and 2015 in 22 countries, covering five continents. This meta-analysis shows that the technique of analysis makes the difference and supervised sentiment analysis seems to best solution to get accurate predictions. At the same time, we highlight that electoral systems matter and the error is lower under proportional representation.

Finally, Chapter 6 opens the way to new streams of research by considering the effect of weighting the forecast to take into account the socio-demographic bias of social media users or by merging social media data with more traditional offline survey polls.

To conclude, drawing a lesson from the 2016 US Presidential elections, the Postscript summarizes and confirms all the main findings reported in this book, remarking the importance of looking at social media as an alternative source of data on public opinion and combining survey data with supervised sentiment analysis to make more accurate electoral forecasts.

References

Ceron, A., Curini, L., and Iacus, S.M. (2016) 'iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content', *Information Sciences*, 367–368: 105–124.

Couper, M. (2013) 'Is the sky falling? New technology, changing media, and the future of surveys', *Survey Research Methods*, 7(3): 145–156.

Daas, P., Roos, M., van de Ven, M., and Neroni, J. (2012) 'Twitter as a potential data source for statistics', Den Haag/Heerlen, the Netherlands: Statistics Netherlands, Discussion paper 201221.

Dalton, R. (2016) 'The potential of "big data" for the cross-national study of political behavior', *International Journal of Sociology*, 46: 1–13.

Groves, R. (2011) 'Three eras of survey research', *Public Opinion Quarterly*, 75(5): 861–871.

Schober, M.F., Pasek, J., Guggenheim, L., Lampe, C., and Conrad, F.G. (2016) 'Social media analyses for social measurement', *Public Opinion Quarterly*, 80(1): 180–211.

1 Social media and electoral forecasts

An overview

Politics, society and social media: a flourishing relationship

The growing number of worldwide citizens with access to Internet, along with the striking growth in their use of social networks like Facebook and Twitter, paves the way to a new potential revolution. In the new "Big Data world", citizens surf the Web, create their own account profiles and share information online, providing scholars with precious data concerning several areas of study that can potentially yield consequences in the real world (King 2014). Social media and social network sites (SNS) generate a large amount of data – particularly (but not only) textual data – which contain information on the tastes and the opinions of social media users. Thanks to the recent advances in text analysis, new quantitative approaches can be applied to these rich sources of data (Iacus 2014; King 2014). As a result, it becomes easier (with some caveats, as we discuss later) to profitably analyze the comments available on the Internet to enhance our understanding of public opinion, paying attention to elements that impinge on the formation of a climate of opinion as well as on the consequences of such opinions.

Not surprisingly, Internet penetration (in 2015, more than 40% of the world population had access to the Web[1] while SNS usage involved 30% of world population[2]), goes hand in hand with the interests of scholars. From the role of Web 1.0 (where Internet is merely one mass medium among the others) to that of Web 2.0, in which social media create a networked, unmediated and interactive online environment, the scientific literature has devoted increasing attention to such new avenues of information. So far, several studies have analyzed the relationship between Internet and social media on one side and politics and society on the other. In this regard, *seven main streams* of research have emerged. These macro-fields try to fulfill a comprehensive assessment of the links between the virtual world of Web and the real world.

The political impact of the Web

A first stream of research looks at the Web as an independent variable and investigates the effect of Internet usage and consumption of online news in order to assess whether, and to what extent, it affects attitudes and behavior of individual users with a particular emphasis on politics (Dimitrova et al. 2014; Stoycheff and Nisbet 2014; Valenzuela et al. 2009). On the one hand, some empirical analyses have attested that the Internet strengthens citizens' demand for democracy (Norris 2011), commitment to democratic governance (Nisbet et al. 2012; Swigger 2013) and satisfaction toward democracy (Bailard 2012). Scholars have found a positive relationship between the use of e-government websites and trust in government (Tolbert and Mossberger 2006; Welch et al. 2005). Moreover, Internet and social media appear to wield positive effects on political knowledge, political participation and several other indicators of civic engagement (Anduiza et al. 2009; Bakker and De Vreese 2011; Boulianne 2009; Jennings and Zeitner 2003; Kaufhold et al. 2010; Kobayashi et al. 2006; Östman 2012; Tolbert and McNeal 2003) while Internet penetration seems even able to increase voter turnout (Miner 2012).

Other studies, however, have underlined a partially different and less optimistic story. Scholars have, in fact, also found a null or negative impact of the Internet on democratic regimes, showing that the Web does not promote either political knowledge and awareness (Groshek and Dimitrova 2011; Kaufhold et al. 2010; Scheufele and Nisbet 2002) or political participation (Quintelier and Vissers 2008), and can even be associated with lower democratic satisfaction (Norris 2011). For instance Avery (2009) did not find differences in trust in government between citizens exposed to online campaign news compared to others; Kaye and Johnson (2002) noticed that the use of the Web for information-seeking purposes is uncorrelated with trust in government; analogously, McNeal et al. (2008) argued that looking for information on institutional websites does not have a statistically significant effect on political trust. On top of that, Im et al. (2014) attested that citizens who spend more time on the Web display a lower degree of trust (although such negative effect is moderated when surfing governmental websites). Falck et al. (2012) and Campante et al. (2013) even witnessed a negative effect of Internet penetration on turnout. What is more, online communities can produce undesirable consequences for the democratic polity as long as they radicalize (rather than moderate) the positions of their users (Alvarez and Hall 2010; Hilbert 2009; Hindman 2009), becoming a source of ideological lock-ins (Sunstein 2001).

Despite these controversial findings on the effects of Web usage, a meta-analysis of Internet studies analyzing 38 different works has shown that Internet is overall beneficial for democracy, even though this positive effect holds only when the Web is expressly used to gather news and retrieve information (Boulianne 2009).

Results of recent cross-sectional analyses of Eurobarometer data related to 27 countries provided more detailed insights that partially confirm such findings (Ceron 2015; Ceron and Memoli 2016). These studies have revealed that Internet usage, per se, has no effect on different measures of democratic support. However, the consumption of online news can make the difference, even though this effect is positive when users consume news from online traditional media, while consumption of news from social media has a negative effect on the satisfaction with democracy (Ceron and Memoli 2016) and on trust in political institutions (Ceron 2015).

Finally, other studies have considered the Web as a "medium" and evaluated how, by providing citizens with additional information, the Web can impinge on the retrospective judgment about the performance of the government (Bailard 2012), and becomes a source to discover government malfeasance and electoral frauds (Reuter and Szakonyi 2015).

Old versus new media

A second stream of literature compares the new media (SNS) with traditional media in order to evaluate whether these latter media (both offline as well as online) keep their agenda-setting power even after the advent of new media or, conversely, if the Web has the potential to exert such "first-level agenda-setting", affecting the attention devoted to certain policy issues by mass media (Hindman 2005; Hosch-Dayican et al. 2013; Jungherr 2014; Meraz 2009; Neuman et al. 2014; Papacharissi and de Fatima Oliveira 2012; Parmelee 2013; Sayre et al. 2010; Vargo 2011; Wallsten 2007).

By reducing the transaction costs typical of old media technologies, SNS leaves room to a new bottom-up style of communication (Benkler 2006), providing egalitarian access to the production and the consumption of news that may no longer be elite biased (Hermida et al. 2014; Woodly 2008) that can potentially break the dominance of traditional media (Lewis 2012; Meraz 2011; Meraz and Papacharissi 2012).

In addition to some anecdotal evidence on their agenda-setting power (for a review: Neuman et al. 2014; Wallsten 2007), scholars have been investigating the impact of blogs and interactive social network sites on the media agenda (McCombs 2005; Meraz 2009) to assess whether they have replaced traditional mass media as a source of first-level agenda-setting, holding the ability to influence the attention devoted to a certain issue of the policy agenda.

In fact, some scholars have provided evidence in this respect (Hindman 2005; Lewis 2009; Lloyd et al. 2006). Woodly (2008) argued that blogs affect the selection and presentation of news stories and set the agenda for political elites like journalists and politicians. Similarly, by analyzing the most popular viral video during the 2008 US presidential electoral campaign, Wallsten (2010) reported that bloggers played a crucial role in attracting media coverage. With respect to SNS, Parmelee (2013) indicated that political tweets play a role in agenda building. Analogously, Conway et al. (2015) investigated agenda-setting effects in seven topics related to the 2012 US presidential primary, showing that in six of them Twitter posts by politicians influenced newspaper coverage. Jiang (2014) found that the Chinese SNS Weibo tends to impact the agenda of the state-controlled television CCTV.

Another stream of literature suggests that social media and traditional media are actually sharing the agenda-setting function but neither is the dominant actor in this process. For instance Meraz (2009) focused on the inter-media agenda-setting dynamics between traditional elite newsroom blogs and top independent political blogs, suggesting that traditional media still retain a strong agenda-setting power, but this power is no longer universal since independent blog platforms redistribute it between mainstream media and citizen media: “traditional media agenda setting is now just one force among many competing influences” (Meraz 2009: 701) and social media seem to enhance citizens’ influence in setting the agenda of media. Analogously, Wallsten (2007: 567) found evidence of a “complex, bidirectional relationship between mainstream media coverage and blog discussion rather than a unidirectional media or blog agenda-setting effect” while YouTube too was found to follow and lead mainstream media salience (Sayre et al. 2010). Along the same vein, Neuman et al. (2014) analyzed Granger causality in a variety of issue areas but found mutual and reciprocal causality between traditional media and social media. Finally, Jungherr (2014) found that Twitter messages follow the same logic of traditional news media in some cases, even though not everything that receives media attention reaches a similar level of attention online.

A wide number of studies, however, have retained that mainstream media still affect the salience of issues discussed online. Traditional media seem to set the agenda of Internet-fueled communication tools, such as bulletin boards and chat rooms (Roberts et al. 2002), weblogs (Lee 2007) and online information seeking (Scharkow and Vogelgesang 2011; Weeks and Southwell 2010). Scharkow and Vogelgesang (2011) resorted to Google Insights for Search to measure the public agenda. They highlighted the potential of such data, suggesting that can even be used in “forecasting tomorrow’s news from today’s recipients’ information seeking” (Scharkow and Vogelgesang 2011: 111). Along this vein, another work has explored the relationship between mainstream media and online search activity, showing that media coverage influences Google Trends’ public salience of a particular topic (Weeks and Southwell 2010). Focusing on Twitter, Vargo (2011) analyzed the reciprocal influence between new and old media, showing that traditional media affect public conversations on a large scale while Twitter has only a limited effect. Analogously, Vargo et al. (2015) analyzed the mortgage and housing crisis agendas and the BP oil spill in the United States, showing that traditional media are still in control of the agenda, even though agenda-setting effects on Twitter are not equal in regards to issues and events.

In this regard, Ceron et al. (2016) recently analyzed the Italian political debate on Twitter and newspapers, focusing on two heated political debates that took place between 2012 and 2014 on issues that were particularly salient for voters, media and the political elites: the reform of public funding of parties, enacted between April and July 2012 after the eruption of several corruption scandals, and the debate over the policy of austerity, which played an important role in view of the 2014 European elections. Their results confirmed that online news media keep their first-level agenda-setting power. The attention devoted to an issue by online media outlets influences the SNS salience of that issue and drives Twitter users to discuss it.

However, there is a difference between being able to affect what the public thinks about (*first-level agenda-setting*) and being able to influence how the public thinks about that issue (*second-level agenda-setting*). In other words, the (possible) enhanced attention built by online media outlets does not imply that they exert a second-level agenda-setting in influencing Internet users (Meraz 2011). For example going back to the previous quoted study, it has found a citizen-elite gap in the degree of anti-politics/anti-austerity sentiment expressed on Twitter, which is remarkably higher if compared to the level of negativity observed in the frame of online news stories (Ceron et al. 2016).

Collective action and public policy

A third stream of literature links social media with collective action and public policy. Under the idea that the Web can decrease the cost of political mobilization, scholars have examined the role of social media in promoting radical protests (Bastos et al. 2015; Calderaro and Kavada 2013; LeFebvre and Armstrong 2016; Segerberg and Bennett 2011) or uprisings (Cottle 2011; Hermida et al. 2014, Howard and Hussain 2011; Meraz and Papacharissi 2012; Shirky 2011; Tufekci and Wilson 2012).

For instance Howard and Hussain (2011) argued that the new digital information technology played a major role in the Arab Spring and social media have changed the tactics of democratization movements. Indeed, with respect to the Egyptian case, Tufekci and Wilson (2012) showed that people learned about the protests primarily through interpersonal communication using Facebook (along with phone contact or face-to-face conversation) and social media use increased a citizen's probability to attend the protests. Analogously, Bennett and Segerberg (2013) discussed the idea of "digitally networked action" and looked at social media primarily as "organizing agents" that can play a crucial role also in offline protests and collective action, as suggested by phenomena such as the Occupy Wall Street movement, or the movement of Indignados in Spain or the global protests against climate change. In this regard, scholars have tried to analyze whether information on online mobilization can be used to make inference on the occurrence of online protests (Bastos et al. 2015). Other studies, however, have downplayed the role of collective digital action and have considered it as mere "clicktivism" or "slacktivism" (Mercea 2017), arguing that online mobilization is a cheap and ineffective but showy form of political engagement (Farrell 2012; Morozov 2011).

Nonetheless in these contrasting findings, online comments are deemed relevant and "dangerous", at least by autocratic leaders who have repeatedly attempted to engage in censorship (King et al. 2013; Shirky 2011) or decided to switch off SNS. Social media comments can also play a role in democratic countries. This can happen either when online mobilization is combined with offline protests (Ceron and Negri 2016) or when there is no relationship between the online climate of opinion and the willingness to organize demonstrations, as shown with respect to the debate on austerity in Italy, France and the UK (Barisione and Ceron 2016). A number of studies has tried to assess whether governments and politicians are responsive to the virtual public sphere, analyzing the effect of online activism on public policy (even though citizen-initiated online participation has been understudied: Dekker and Bekkers 2015: 9). Some studies have shown examples of governments' responsiveness to the virtual public sphere; others have found that online participation is largely ignored. For instance by analyzing the impact of the pressure put by social media users on their elected officials during the selection of the Italian Head of State in 2013, we notice that – at least with respect to such key strategic institutional choices – the ability of SNS to affect democratic political systems by facilitating the interaction between citizens and voters appears substantially limited, even in this favorable context where SNS allowed citizens to voice and exert substantial pressure (Ceron 2016a). In such case, the choice of MPs is still affected by the traditional "competing principals" that drive legislative behavior, such as the party leadership, the factional affiliation or the constituency of voters. In the field of public policy, however, a meta-analytical review of existing studies (Dekker and Bekkers 2015) has found that responsiveness can sometimes occur, depending on the policy domain object of study and on institutional characteristics such as features of the policymaker and of online participation.

Finally, other studies have focused on SNS to try to extract from that the opinion of the social media users on topics such as anti-American attitudes and support for terrorist attacks (Ceron et al. 2015; Jamal et al. 2015; Malik 2014; Zeitzoff et al. 2015) or Euroscepticism and support for the European Union and its policies (Barisione and Ceron 2016; De Wilde et al. 2014).

E-campaigning

Two other streams of research on SNS and politics deal more directly with elections, political parties and policy platforms. Recent studies have started to investigate the phenomenon of e-campaigning, that is the forms through which political parties or individual candidates campaign online before an election (Hosch-Dayican et al. 2013; Vergeer et al. 2013). Scholars have scrutinized how the Internet impacts the content of electoral campaigning (Ceron and Curini 2016; Druckman et al. 2010; Evans et al. 2014; Lev-On 2011) to evaluate whether patterns of e-campaigning are similar or different compared to traditional offline campaigning (Gibson and McAllister 2014). Some studies have suggested that e-campaigning follows different patterns when compared to traditional campaigning, while others support the "normalization thesis", arguing that there is no difference between the political parties' usage of the Web or traditional media (e.g. Schweitzer 2008). At least in the era of Web 1.0, it seemed that online campaigns spread only the same content that was broadcast offline (Druckman et al. 2010; Ku et al. 2003). Things might have changed after the rise of social network sites in the interactive Web 2.0, where users can more easily generate their own content through social media, and candidates can potentially develop personalized and targeted messages, different from the party's official campaign (e.g. Vergeer et al. 2013). The debate is still open (Gibson and McAllister 2014), as some works have attested a change (Strandberg 2013) while others have provided support for the normalization thesis (Vergeer and Hermans 2013), even in contexts where intra-party competition should provide candidates with the incentive to enact tailored e-campaigns (Vergeer et al. 2013). If the Internet is "nothing more than an extended tool to distribute the same information used in offline campaigning" (Vergeer et al. 2013: 482), the analysis of the effects of e-campaigning can also allow us to catch the dynamics of the whole campaign (Druckman et al. 2010; Evans et al. 2014) and evaluate the effectiveness of alternative strategies. The fact that political activities on social media mirror offline political action just strengthens the previous point: for example MacWilliams (2015) showed how the relative effectiveness of campaigns in enlisting, engaging and mobilizing Facebook users can be theorized as a proxy for estimating the effectiveness of a campaign to generate support, activism and votes among voters.

In light of this, some works also have evaluated the effectiveness of different e-campaigning strategies in terms of electoral success, assessing whether the usage of social media tools for e-campaigning allows a candidate to win more votes in countries like Australia, Denmark, Ireland, Sweden or the Netherlands (Gibson and McAllister 2014; Hansen and Kosiara-Pedersen 2014; Jacobs and Spierings 2016; Larsson 2015; Sudulich and Wall 2010) and to disentangle whether they can be electorally rewarding on the basis of alternative contents broadcast online (Ceron and d'Adda 2015). These studies have shown that, although negative e-campaigning seems more effective in winning votes (Ceron and d'Adda 2015), the official social media accounts of political parties mainly resort to positive campaigning rather than negative campaigning (Ceron and Curini 2016; Evans et al. 2014; Hosch-Dayican et al. 2016). We discuss at length about this last point in Chapter 4.

Estimating policy positions

Taking the cue from the idea that parties and politicians broadcast online their political views and express their tastes, both before and after an election, other scholars have tried to estimate the policy positions of political actors (and citizens: Barberá 2015; Bond and Messing 2015) by analyzing social media data (Barberá et al. 2015; Boireau 2014; Boutet et al. 2013; Ceron 2016b; Conover et al. 2011; Ecker 2015; Hanretty 2011; King et al. 2011). These studies have relied on three different approaches.

The first approach is based on the communication patterns that take place on social media (Boutet et al. 2013; Conover et al. 2011). These studies have analyzed elements such as the number of political messages sent by political elites and pay attention to who replies to such messages and to who shares and forwards that content. On the basis of these data, scholars have tried to identify different patterns of behavior that are analyzed in order to locate individuals and political actors on an ideological scale. Such approach has been successfully applied, for instance to classify Twitter users during the US midterm elections in 2010 (Conover et al. 2011) and the UK general elections in 2010 (Boutet et al. 2013).

A second alternative looks at the structure of online social networking sites. In detail, scholars have analyzed the network structure of voters and political elites to retrieve estimates of individual policy positions (Barberá 2015; Barberá et al. 2015; Bond and Messing 2015; Ecker 2015; Hanretty 2011; King et al. 2011;

Livne et al. 2011). This method adopts a two-step process: in the first stage, the network is transformed in a voter-to-elite adjacency matrix. Then, by applying data reduction techniques, this approach estimates individual policy positions on an ideological latent dimension.

Finally, a third method focuses on the actual content of messages sent by political actors from their social media accounts. By identifying and collecting all relevant messages (under the idea that they express a political content), scholars can detect different patterns of word usage that are adopted by political actors to present their ideology and their policy preferences. By comparing word usage using text analysis techniques, it becomes possible to place political actors on a latent dimension (Boireau 2014; Ceron 2016b; Livne et al. 2011; Sylwester and Purver 2015).

These different techniques look promising and can, indeed, open new avenues to political research, as they provide new tools to monitor individual preferences and to track, potentially in real time, the evolution of the opinions of political actors matching such opinions with their actual behavior.

Emotions and well-being

The sixth research stream looks at SNS as a source of information to measure the attitudes of the online public opinion. From this perspective, recent studies have started to analyze the content of the opinions expressed online to investigate sensitive topics so different among themselves such as happiness (Curini et al. 2015), racism and intolerance (Burnap and Williams 2015; Stephens 2013). For example Stephens (2013) exploited Twitter to create a "hate map" that identifies the degree of racism and intolerance geo-located across the United States. This tool can be useful to adjust the educational policy and to prevent episodes of interracial violence. Burnap and Williams (2015) provided a more in-depth analysis of online hate speeches. They combined a supervised machine learning text classifier with an ensemble meta-classifier to foresee the spread of cyber hate on Twitter. Curini et al. (2015), instead, analyzed Italian tweets to monitor the daily level of happiness, showing that it is affected by not only meteorological elements but also the spread between German and Italian Bonds. Other studies have tried to predict the evolution of public mood, focusing on the level of stress as well as on anger and joy (Lansdall-Welfare et al. 2012).

From this perspective, a new area of research is related to studying subjective well-being by analyzing perceptions expressed on social media. In fact, the literature on well-being measurement seems to suggest that, while "asking" for a self-evaluation is the only way to estimate a complete and reliable measure of well-being, at the same time "not asking" is the only way to avoid biased evaluations due to self-reporting (Curti et al. 2015; Iacus et al. 2015). Just to mention a few more studies: Schwartz et al. (2013) examined tweets from 1,300 different US counties, measuring life satisfaction through the recurrence of words used and their topics, defined as sets of recurring words referring to a specific domain (e.g. *outdoor activities*, *feelings*, *engagement*). This language analysis was found to be predictive of the subjective well-being of people living in those counties as measured by representative surveys, and the topics identified provided more detailed behavioral and conceptual insights on counties' life satisfaction than the usual socio-economic and demographic variables. Finally, Durahim and Coşkun (2015) compared "official" statistics of well-being (measured with surveys on a province basis by the Turkish Statistical Institute) with the results of a social media analysis (led on 35 million tweets published in 2013 and in the first quarter of 2014) and found high similarity among them.

Economic, social and political forecasts

The last stream of research connecting SNS and politics deals with forecasts, under the idea that information available online can be useful to anticipate dynamics and predict trends (Leetaru 2011; Weinberger 2011). For example the US government has tried to use digital data and digital texts made available on social media to predict the occurrence of socio-political events. In this regard, two projects have been developed. One is called Open Source Indicators (OSI). It analyzes data available online not only to monitor the flow of ideas, goods and people but also to predict the changing sentiment of citizens and to detect, for instance anger and pessimism, which in turn can be associated with the occurrence of economic crises or revolts (Weinberger 2011). Along the same vein, another famous project is called Recorded Future. It has been promoted by Google and the CIA to analyze websites, blogs and Twitter accounts in order to find relationship between people, organizations and events (Helbing 2013). Other academic analyses have tried to anticipate revolts or coups by analyzing the *New York Times*' archives of news (Radinsky and Horvitz 2013) and open source data retrieved from Wikipedia or other online sources; similarly, Leetaru (2011) showed that the analysis of news and social media comments was useful to predict the revolts of the Arab Spring in Tunisia, Libya and Egypt (see also Koehler-Derrick and Goldstein 2011 on the Egyptian revolution), as well as the stability of the Saudi regime, or to detect the refuge of Osama Bin Laden (predicted within a 200 km range).

With respect to economics, several studies have put the "sentiment" (i.e. the mood) of people in relation with the trend of stock exchange. For instance Antweiler and Frank (2004) showed that the volume of comments published on selected finance forums is a good predictor of the volatility of stock markets. Other studies have analyzed blogs (Gilbert and Karahalios 2010) or tweets (Bollen et al. 2011), showing that the mood of social media allows us to forecast variations in the Dow Jones index (Zhang et al. 2011) as well as the value of gold, oil and the exchange rates of currencies (Zhang et al. 2012). Analogously, other scholars have disclosed relationships between the changes in the volume of queries related to economic topics (measured using Google Trends) and the Dow Jones index (Preis et al. 2013). Google queries have also been put in relation with labor market data on unemployment (McLaren and Shanbhogue 2011). The same authors managed to predict the growth of the real estate market by analyzing online data on estate agents; finally, they put Google search queries on Value-Added Tax (VAT) in relation with data on the Consumer Confidence Index (CCI), even though in this latter case no effect was found (McLaren and Shanbhogue 2011).

The predictive power of the Web has also been extensively tested in the areas of marketing, communication and show business. Several studies have analyzed comments published on blogs and social media to predict sales (Liviu 2011), focusing on books (Gruhl et al. 2005) or movies (Huberman and Asur 2011). The idea is that posts not only reflect individuals' own thoughts but also express judgments about the larger social world. Analogously, social media data – particularly tweets – have been used to predict the ranking and the winner of television shows in which the outcome is based on the preferences of the audience: in this regard, reality shows like *American Idol* (Ciulla et al. 2012) have been analyzed. The "wisdom of the crowd" has been used also to predict contests in which social media users do not influence the outcome. For instance many studies have tried to predict the winners of the Academy Awards (Bothos et al. 2010; Liviu 2011), while the results of the Football World Cup 2010 have been predicted too (Uzzaman et al. 2012), and indeed, the output Twitter analyses was somewhat in line with the previsions of bookmakers.

One last avenue of social media-based forecasts involves prevention, in particular with respect to the diffusion of earthquakes and the damages they caused (Sakaki et al. 2013), as well as to public health concerns. In 2009, for example, Google Flu Trends started to analyze search queries to predict the occurrence of an epidemic contagion (Freifeld et al. 2008). Scholars have shown that the number of queries was correlated with the frequency of flu symptoms in the United States (Ginsberg et al. 2009), and these results were confirmed also when analyzing European countries (Achrekar et al. 2012; Valdivia et al. 2010), although the controversy on such results has not been lacking (Butler 2013; Cook et al. 2011). Along this vein, Signorini et al. (2011) used the content embedded in the Twitter stream to track public sentiment on H1N1 virus, or swine flu, building a measure of "public concern" about health-related events, while, Lampos and Cristianini (2012) predicted the diffusion of the flu in the United Kingdom using Twitter data.

Many studies involving social media data are related to drug abuse too. For instance Murphy et al. (2011) analyzed online conversations about drug use, showing that these are correlated with actual offline rates of drug usage. Similarly, Hanson et al. (2013) analyzed Twitter comments focusing on the mentions of "Adderall"; they detected prescription drug abuse in the US, which showed up during traditional college final exam periods, particularly in the country's college regions. Finally, Cunningham (2012) investigated the temporal trends in the consumption of alcohol and cigarettes across the hours of the day and the days of the week and correlated these trends with the actual offline behavior.

Social media and electoral forecasts: a crowded crossroad

The idea of digging into the Web in order to predict socio-political outcomes can immediately be extended to the field of electoral forecasts. The rest of the book is devoted precisely to this. To build a predictive model and validate its predictive accuracy, we need to fulfill three requirements (Wu 2012)[3]: 1) we need a model or algorithm that computes some predicted outcome (e.g. stock price of a company, weather in San Francisco tomorrow, etc.), 2) we need to have an independent measure of the observed outcome that the model is trying to predict, and 3) we need a measure that compares and quantifies how closely the predicted outcome matches the independently measured outcome. The most difficult to get of these requirements is the second, that is having an independent measure of the outcome.

In this regard, forecasting an election is one of the few exercises on social events where a clear and indisputable measure of the outcome that a model is trying to predict is available. Furthermore, such a measure, which is the vote share of candidates (and/or parties) at the ballots, is completely objective and comes from an independent source.

The first attempts to exploit information available online to predict the results of elections date back to 2007, as we discuss in the section on the computational approach to social media-based electoral forecasts. The literature on social media-based electoral predictions, therefore, is still in its infancy, even though many studies have been performed and many (different) attempts have been made.

The variety and the originality of such attempts depict a new and vibrant field of research, which lies at the crossroad between computer science, political communication, linguistics and political methodology. As anyone can guess, at a crossroad like this there is a lot of chaos.

This field of research is developing at a fast pace, and many alternative techniques (sometimes very similar to, sometimes very different from each other) have started to be adopted by scholars from all over the world. Most of the time, scholars have used their own technique to forecast a single election in a single country. As such, it becomes difficult to stay updated on the advances of social media-based predictions, and scholars can easily lose their bearings. Just to quote an example, in 2014, a study claimed to have followed and improved on the state of the art in the analysis of Twitter data by analyzing mentions of parties and candidates (Caldarelli et al. 2014), while the study completely ignored other streams of research based on the content and the sentiment of Twitter comments rather than the number of mentions only.

Such confusion due to the variety of approaches used and fields involved, which also makes it hard to have a full knowledge of the literature, becomes problematic and represents a hurdle to establishing a collaborative environment bringing researchers together to improve on existing studies.

Beside using different techniques, in the attempt to produce electoral forecasts, scholars have also analyzed a huge variety of Internet sources, ranging from blogs and social network sites like Facebook (Barclay et al. 2015; Cameron et al. 2013; Giglietto 2012; Lindsay 2008) and Twitter (Ceron et al. 2014; Gayo-Avello 2012; Tumasjan et al. 2010), to online news (Fonseca 2011; Véronis 2007), Google search (Reilly et al. 2012; Whyte 2015) and Wikipedia page view data (Yasseri and Bright 2015), while some scholars have also tried to mix data coming from different sources (Franch 2013).

As an attempt to bring some order into this growing literature, we first present and discuss the main alternative approaches toward predicting elections through

Table 1.1 A typology of methods employed to perform social media-based electoral forecast

Main approach	*Sub-approaches*
Computational	Volume data Endorsement data
Sentiment Analysis (SA)	Traditional Sentiment Analysis (ontological dictionaries, NLP) Machine Learning
Supervised Aggregated Sentiment Analysis (SASA)	ReadMe iSA

social media. We then contrast such approaches to see the advantages and the pitfalls of each as well as the logic behind them.

According to the method adopted to perform a prediction, the existing studies can be classified in three main approaches, which in turn can be split into subcategories (see Table 1.1). A first approach is merely quantitative and relies on automated computational techniques to count data. A second approach pays attention to the language and tries to attach a qualitative meaning to the comments (posts, tweets) published by social media users, employing tools for sentiment analysis. A third method follows a similar perspective though performing supervised semi-automated sentiment analysis to catch the (aggregate) opinion expressed online. Each is discussed in greater detail in the next sections.

The computational approach to social media-based electoral forecasts

The computational approach to social media forecasts focuses on using an automated technique to enumerate quantity of interests, whose mere count is considered informative of the potential outcome of an election. The roots of the computational approach rely on the analysis of political coverage of candidates provided by mass media. The very first attempt in predicting elections, in fact, was made by looking at news. Véronis (2007) analyzed the first round of French presidential elections in 2007. He measured the number of times each presidential candidate was cited in 2,200 political articles found through the Rich Site Summary (RSS) feed, published during the last week of the campaign, showing that the citations on the press were a quite good predictor of the electoral results, with an accuracy rate higher than the last survey polls made available 2 days before the election. This surprising result can be effectively considered the starting point of a stream of research. Taking the cue from such attempt, a few years later, other authors tried to analyze online news in order to forecast local and national elections in Japan (Senkyo Kakaricho 2009, see also http://senkyo.kakaricho.jp/) and Portugal (Fonseca 2011).

At first sight, the rationale behind the use of news citations as a source of information on electoral outcomes seems crystal clear. The media are more likely to pay attention to politicians who have a real chance to win the elections, and, therefore, potential successfully candidates are cited more in the press. When the race is tied, for instance, mass media mainly talk about the two leading candidates compared to a situation in which third parties can have a chance to compete (getting access for example to the runoff). Actually, this reasoning implies either that mass media rely on information provided by survey polls or that the hearts and the minds of political journalists are in line with those of the whole public opinion. In the first case, media become just the megaphone of survey polls, which can be problematic when surveys are not available; this is something that does occur in some countries in the last days before the election, that is in those days in which voters make their choice and surveys could become a more credible and powerful predictor of results. For example in Italy the law specifically forbids the publishing of polls in the 2-week period preceding an election, that is polls can be conducted in this 2-week period, but their results cannot be rendered public (see Callegaro and Gasperoni 2008). Analogous restrictions to the publications of preelectoral surveys are in place also in other countries such as Bulgaria, Czech Republic, Peru and Singapore; restrictions exist also in France, Canada, Russia and Albania, though for a shorter amount of time.[4]

Conversely, the assumption behind the second case can be more questionable in political systems in which an ideological media bias exists and there is a detachment between the viewpoints of journalists and voters (Curini and Splendore 2016; Puglisi and Snyder 2016).

Following the idea of recording the attention toward rival candidates, later studies have tried to cope with this concern. Successful parties and politicians, in fact, should attract more attention than others not only within the newsrooms but also in the public conversations of common citizens. Going beyond the number of citations in the press, then, scholars have measured the degree of attention devoted to candidates by users of SNS (DiGrazia et al. 2013; O'Connor et al. 2010). While the first approach (looking at media coverage) is merely based on the judgments of "experts" like media journalists or, indirectly, pollsters, the second approach resembles more the logic of the "wisdom of the crowds" (Franch 2013; Surowiecki 2004).

The wisdom of the crowds is the idea that, in the aggregate, the collective opinion of a group of individuals is more suitable to produce an accurate estimate of a quantity of interest (by taking the average of the individuals' estimates) rather than the single opinion of any member of the group. Following that, it can be argued that the crowd is wiser than the experts. This should be particularly true in the case of SNS. Indeed, to be wise, the crowd needs to be diverse and independent and its decisional procedure has to be decentralized (Surowiecki 2004). This is something that perfectly applies to social networks in the Big Data world.

The "social media crowd" also has another peculiar trait. As long as we live in a hybrid (and networked) media environment (Chadwick 2013), the news media agenda interacts with the judgments and the reaction of online media users, who contribute to the production of news and to their diffusion. It has been argued (see our earlier discussion) that mainstream media maintains their agenda-setting power (Ceron et al. 2016) up to the point that, for instance media events like political talk shows or debates between candidates broadcast on television are drivers of citizens' conversations on Twitter (Jungherr 2014). Even so, the social media crowd does not merely reflect political media coverage but rather it is able to evaluate which type of coverage deserves better attention as only some kind of news goes viral.

Similarly, in the last days before Election Day, social media users do pay attention to the political debate. For the reasons just discussed, such information could be a better predictor of electoral outcomes if compared to the news. The data generating process behind such kind of analyses is not much different from that behind other attempts made to produce forecasts on the winners of television shows, song contests or football matches that we have mentioned earlier.

However, social media users are not only going to discuss which candidate will win. In fact, they can most easily discuss around which party or which candidate to support. Eventually, they can also make a choice and declare it online. This is exactly what happened during the campaign for the 2012 US presidential campaign, when the Pew Research Center reported that no less than 22% of voters spontaneously declared their voting behavior on social network sites like Facebook and Twitter (Pew Research Center, 2012). With numbers like these, we can easily understand why many scholars have started to consider comments published online as an expression of support for parties and candidates, rather than a mere indicator of which candidate is more likely to win.

As such, early studies simply have assumed that the number of online posts *mentioning* each party mirrors its degree of offline approval. One of the first and most well-known attempts, in this regard, is the study produced by Tumasjan et al. (2010) related to the 2009 German federal elections. The authors argued that the content of tweets is informative of the political sentiment of voters. Accordingly, they disclosed the predictive power of social media and claimed to have predicted the results of the election, reporting, on average, a 1.65% deviation between the share of Twitter mentions and the actual share of votes won by the main German parties. This result was astonishing since the authors showed that the average error was in line with that of traditional survey polls.

Despite this success, such approach has been strongly criticized (Jungherr et al. 2012; Jungherr et al. 2016) because the number of mentions received by a party/candidate, *per se*, may be not informative on the degree support retained by this party/candidate within the population of (Internet) voters. In particular, many comments mention more than one candidate, for instance because the author often wants to express support for his or her preferred candidate while attacking the rival one. Focusing on the number of mentions alone could then be problematic any time that we find multiple names and, in fact, such type of texts are often discarded by researchers, producing a loss of substantive information that may bias the results, particularly if the propensity to criticize is not equally spread among supports or rival candidates. This is something that can happen, for instance with respect to the front-runner candidate as the literature on negative campaigning suggests.

Other attempts have, therefore, tried to refine the count of mentions by measuring only some peculiar citations or selected hashtags (Cosenza 2013; Cunha et al. 2014; Jungherr et al. 2012, 2014). For instance in the German case (Jungherr 2013), scholars took advantage of the fact that, in 2009, voters used to post tweets that included the hashtag of a party's name followed by a plus or a minus (e.g. *#cdu+* or *#cdu–*). Such a hashtag was then interpreted as an expression of support or aversion toward the party (and, in some cases, this attribute was erroneously considered equal to the "sentiment", see following). In the Italian case, another attempt was made in order to count the number of mentions of a party name joint with a statement like "I will vote for . . ." or "I stand with . . ." (Cosenza 2013), while in Brazil Cunha et al. (2014) some peculiar hashtags related to the official campaign of candidates were recorded.

Whether focused on the mentions on online news or SNS, all these approaches share some common traits and they fall into the category of "*volume data*" (see Table 1.1).

Other computational techniques, however, do not rely on the volume of mentions but consider other indicators. Some studies have counted the number of Twitter followers or Facebook friends, while others have paid attention to that the number of likes received on the Facebook walls. All these attempts can be summarized under the label of "*endorsement data*" (see Table 1.1). Williams and Gulati (2008) were the first to show that the number of Facebook supporters could be considered as a valid indicator of electoral success. From this perspective, the choice of becoming a friend or a follower of a politician can be interpreted as an intention to support him or her. Nevertheless, such a proxy is not able to take into considerations the possibility that users can choose to follow a famous politician just to stay updated on his or her ideas and behaviors. Indeed, Twitter users are often divided between those who follow leaders they agree with and those who also follow political figures with which they disagree (Parmelee and Bichard 2011).

For this reason, other studies have taken into account the number of likes recorded on the Facebook pages of parties and candidates (e.g. Barclay et al. 2015; Cameron et al. 2013; Giglietto 2012). Barclay et al. (2015) found that in India the number of likes a party or its leader secured on their official Facebook fan page was strongly and positively correlated with the actual vote share. Conversely, in the 2011 Italian mayoral elections, the correlation between likes and votes was not significant, even though the most popular candidate on Facebook actually won the race 4 times out of 10 (Giglietto 2012). From this perspective, a more comprehensive attempt was made by Franch (2013) in the context of the 2010 UK election. He analyzed such information from multiple social media sites (Facebook, Twitter, Google and YouTube), showing that computational data were useful and informative when combined with the estimated produced by traditional survey polls.

Arguably, the number of followers or friends remains just a measure of awareness; the number of likes too can rather represent short-term attention more than actual support as suggested by Cameron et al. (2013). Furthermore, the number of likes is also related to the number of friends/followers, as you can easily get more likes when your audience is larger. Both the number of friends/followers and the number of likes (or its variation) then can be misleading.

This aspect appeared evident, just to report a well-known example, in the 2012 US presidential election: at the beginning of the campaign, Barack Obama, who was the incumbent president, retained almost 17 million followers on Twitter, while his opponent, Mitt Romney, had far less than 1 million. Despite this, the race between the two candidates was very close, particularly if we look at the popular vote (as we discuss in Chapter 3 more in detail).

These predictions have been also criticized for a number of other reasons (Gayo-Avello 2013). In particular, it has been argued that the performance of prediction based on the volume of data is too unstable, primarily because these methods are way more dependent on arbitrary choices made by the researchers about what parties and candidates to include in the analysis or about the lapse of time of data gathering (e.g. Jungherr et al. 2012, 2014) and the selection of keywords or hashtags used in data collection. For these reasons, it would be hard to attest that successful predictions have not occurred just by chance or as an artifact of such arbitrary choices tailored on the estimates of survey polls or on the final results of the election.

Summing up, no matter whether they record attention, awareness or support, merely computational data seem to retain some problematic attributes and might fail to catch the informational complexity of the social media environment.

Sentiment analysis and machine learning applied to electoral forecasts

In the previous section, we noted that as Election Day approaches, social media users tend to discuss more about political topics; in their conversation, users can also take a side and express support for one of the alternative political options. Not all the comments toward a candidate, however, are necessarily positive. As the literature on negative campaigning has suggested (Hansen and Pedersen 2008; Skaperdas and Grofman 1995), before an election we can also observe a huge number of negative statements made to criticize the opponent candidate rather than to support the favorite one. This feature also applies to the social media environment, and negative messages are broadcast by politicians, parties, candidates and official accounts of political actors (Ceron and d'Adda 2015; Evans et al. 2014) as well as by common citizens and voters (Hosch-Dayican et al. 2016). For instance a study on the communication strategies enacted by citizens during the 2012 Dutch election campaign has shown that citizens participate significantly in online electoral campaigning on Twitter, and they are more prone to use negative campaigning and to express discontent, compared to the communication strategy of the political elite (Hosch-Dayican et al. 2016).

These results clearly highlighted that candidates can be mentioned not only by those who are willing to vote for them but also by voters who want to criticize them. For these reasons, going beyond computational approaches, other studies have adopted various techniques of sentiment analysis (SA) to supplement data based on the volume of mentions, where for SA we mean, at least for now, analyzing texts to extract information (see Chapter 2 for a more detailed discussion in this respect). Some of these studies have focused on the net sentiment, that is the difference between the rate of positive and negative messages linked to a candidate. Others have kept a focus on the extent of online support (in line with the rationale discussed earlier), and they pay attention only to the degree of positive sentiment.

Arguably, the idea to apply sentiment analysis techniques to forecast electoral results is almost as old as the count of citations of a candidate. The very first attempts date back to the 2008 US presidential election, when Lindsay (2008) developed a sentiment classifier based on lexical induction and found correlations between polls and the content of Facebook wall posts. Another well-known study, which has shed light on the potential of text analysis, has confirmed that the sentiment expressed by Twitter users in favor of Obama is roughly correlated with his degree of support recorded by multiple by survey polls during the 2008 electoral campaign, but this study has found a stronger correlation between Twitter sentiment and Obama's approval rate measured by Gallup in the aftermath of election, that is in the course of 2009 (O'Connor et al. 2010).

These first results stimulated others to test the predictive power of sentiment analysis, and during the following years, many attempts were made by scholars in this field. They analyzed the sentiment (particularly Twitter sentiment) in a variety of countries and elections like Australia (Unankard et al. 2014), France (de Voogd et al. 2012), Greece (Tsakalidis et al. 2015), India (Khatua et al. 2015), Ireland (Bermingham and Smeaton 2011), Nigeria (Fink et al. 2013), Singapore (Choy et al. 2011), the Netherlands (Sanders and van den Bosch 2013; Tjong Kim Sang and Bos 2012), the United Kingdom (Lampos 2012) and the United States (Choy et al. 2012; Washington et al. 2013).

Most of these predictions reached relatively low average errors, ranging from around 5 points up to almost 0, with the notable exception of Nigeria, where the average error of the prediction was no lower than 9.

Overall, the results were promising. For instance, de Voogd et al. (2012) examined Twitter in light of the 2012 French presidential campaign and found high correlations between media exposure, public opinion polls and sentiment measures of contents of Tweets, while Tjong Kim Sang and Bos (2012) showed that SA performed as well as polls in predicting the results of the 2011 Dutch senate election.

Nevertheless, not all the automated techniques of SA are equal. Indeed, all of these attempts were made by using different tools provided by several private commercial companies or built ad hoc by the researchers. Although they may share some similarities, each of these techniques also has some peculiarities which differentiate one from the others. To allow for a comparison, we focus on two broad categories and we roughly distinguish between *traditional sentiment analysis* (i.e. those techniques based on ontological dictionaries, corpora, lexicon-based classifiers and other natural language processing and mostly unsupervised) and *machine learning* (i.e. those techniques that perform individual prediction of the semantic content of texts starting from a given training set of data). The ontological dictionaries or semantic orientation techniques involve the application of a previously constructed word set containing positive and negative terms, and thus, the automated identification of polarity is determined using the frequency of these terms in the text, without the need for a sample database that has previously been categorized (what is also called unsupervised learning) (Ghiassi et al. 2013; Olivera et al. 2015; Yu et al. 2013). The machine-learning techniques, instead, require the manual sorting of database samples (through the selection of a training set) to build models based on text that serve as parameters in automated data analysis (what is also called supervised learning) (Guerra et al. 2011).In the first category we can mention, for example O'Connor et al. (2010), who relied on existing lexicon-based dictionaries composed of a list of terms labeled as positive or negative or the already quoted Lindsay (2008), who developed a sentiment classifier based on lexical induction. Among the *machine learning* classifiers we can mention the work by Bermingham and Smeaton (2011) on the Irish elections and by Tjong Kim Sang and Bos (2012) related to the Netherlands. We discuss more in depth these and other techniques in Chapter 2.

Limits and criticism of social media-based electoral forecasts

Despite promising preliminary results, the very same possibility to develop social media-based electoral forecasts has been highly criticized. Some of the weaknesses highlighted in the existing literature (that add to the ones already discussed) are also common to computational approaches, while others are more peculiar to sentiment analysis.

We start from the more general remarks. First, many of the predictions published so far were not real ex ante predictions but rather post hoc analyses. While many studies claim to have predicted the outcome of elections using social media data, Gayo-Avello (2013) correctly noticed that most of those studies were not predictions at all. In fact, several analyses have been performed only after the election and were made publicly available only post facto (Gayo-Avello 2013). As such, they cannot be defined as "predictions", and we should better refer to them as "post-dictions". By definition, in fact, electoral forecasts must be made and published before the elections take place (Gayo-Avello 2013). The traditional literature on electoral forecasting openly recognizes that "the forecast must be made before the event. The farther in advance [. . .] the better" (Lewis-Beck 2005: 151).

In view of that, it becomes hard to gauge whether successful estimates are not due to a bias of the researcher, whose arbitrary choices on the design of the analysis could have been affected by knowledge of the final results. Along the same vein, the fact that many predictions have been published only ex post points to the risk of having a selection bias in the publications of the results. Scholars, in fact, can be more prone to public (on a blog, an online repository or an academic journal) positive results (successful predictions) rather than unsuccessful ones (Franco et al. 2014). In other words, the fact that we tend to observe a link between volume data or sentiment and electoral results could be the artifact of such selection bias, as scholars who have produced unsuccessful predictions may be willing to hide them or give up from trying to publish them.

This criticism is strongly related to another one. Almost each study has developed its own tool for sentiment analysis or machine learning and its own technique on how to count mentions (e.g. with regard to the selection of keywords or hashtags used in data collection or on the time span of the analysis). The very same technique usually has been adopted to predict only a single election or elections in a single country. Therefore, doubts arise about how confident we can be on the generalizability of the results of such single case studies. This point is even more prominent as only few exceptions have tried to adopt a comparative perspective using a similar technique on a variety of cases across countries (e.g. Ceron et al. 2014; Choy et al. 2011, 2012; Gaurav et al. 2013; Nooralahzadeh et al. 2013; Tsakalidis et al. 2015).

A third general criticism concerns that fact that the accuracy of the prediction should be evaluated against a clear and coherent indicator of success. Scholars have compared the prediction either against survey polls or against the outcomes of the election. Even so, when scholars have compared their prediction with the electoral results, they have used very different indicators to evaluate the performance. For instance some scholars have focused on whether the forecast manages to predict the winner of the race. Here the choice of the baseline model is problematic too, as scholars can claim to have successfully foreseen the outcome even when their model is less powerful if compared to the traditional model used in the political science world and based on the idea of incumbency advantage.

Conversely, others studies have looked at the performance by recording the correlation between the prediction and the actual share of votes or by looking at the level of confidence of the coefficient of the regression in which the actual result is the dependent variable, while the prediction is one of the independent variables. Most of the prediction, however, has relied on the mean absolute error (MAE). The MAE is measured as the average of the absolute difference between the prediction and the actual result (or between the prediction and the baseline adopted for comparison). The MAE has been widely used to compare the accuracy of forecast based on social network analysis (Tumasjan et al. 2010) and that of political information markets relative to election polls (Huber and Hauser 2005). Although this indicator is not perfect (its limitations are discussed in Beauchamp 2014; Gayo-Avello 2013), it is particularly useful to compare the accuracy of a wide number of forecasts precisely due to its widespread diffusion. Despite the availability of alternative measures of accuracy that are sometimes used in the literature, such as the R^2, the standard error of estimate, the root mean squared error (Lewis-Beck 2005; Martin et al. 2005) or more complex measures of predictive accuracy (see Arzheimer and Evans 2014), in many cases the MAE is the only indicator made available by the authors of the forecast (see Chapter 5).

Two additional sources of skepticism toward social media-based electoral forecasts are related to the individual traits of social media users.

On the one hand, social media users are not necessarily representative of the socio-demographic traits of the electorate (Bakker and De Vreese 2011; Vaccari et al. 2013). They tend to be young and highly educated males, even though these differences seem lower when we focus only on people who express political opinions (Bakker and De Vreese 2011) or when we consider their auto-collocation on the left–right scale (Vaccari et al. 2013). Nevertheless, any potential mismatch between the traits of social media users and that of the whole population has largely been ignored by many studies, even though few attempts have been made in order to correct the forecast by considering the socio-demographic features of users or other external type of information which was used to apply some forms of weighting (Choy et al. 2011, 2012; Lampos et al. 2013; Shi et al. 2012).

On the other hand, social media are also affected by a self-selection bias, which has largely been ignored in the literature (Gayo-Avello 2013). When dealing with unsolicited opinions, scholars have to recognize that the comments will reflect more the opinion of those more active (and politically active). A limited number of users can talk a lot and can repeatedly express opinions, while the wide majority of accounts is quieter and barely sends a message (Mustafaraj et al. 2011). This can be true for individuals but also for collective groups and some "vocal minorities" who can express their voice louder than a silent majority.

The last criticism highlighted in the literature, which is probably the most relevant limitation directly addressed against sentiment analysis, is related to its naïveté (Gayo-Avello 2013). Scholars have recognized that traditional SA can be a noisy instrument (only slightly better than random classifiers) in the field of electoral prediction because it is particularly difficult to deal with the political language used on social media (Gayo-Avello 2011, 2013). The main problem is that unsupervised sentiment analysis techniques are generally not well suited to handle humor, double meanings and sarcasm and, therefore, may not accurately catch the actual meaning of the opinions expressed, particularly when dealing with short sentences in which a single word can completely alter the meaning of the statement (Gayo-Avello 2012; Jungherr et al. 2012; Metaxas et al. 2011). An additional limit of automated/unsupervised SA approach is that it relies on predefined ontological dictionaries and thus cannot address the risk of a spamming effect due to the presence of noise and misleading information or changes in the language.

The majority of the discussed concerns can (at least partly) be addressed by relying on a proper methodology for sentiment analysis. Indeed, recent advances in quantitative text analysis allow researchers to integrate quantitative large-N data with in-depth analyses and reopen the debate around whether analysis of social media can be helpful in predicting election results. Beside this, other studies have started to adopt statistical approaches mixing social media data with the estimates of traditional opinion polls in order to enhance accuracy of the estimates (Beauchamp 2014; Franch 2013; Tsakalidis et al. 2015).

Accordingly, these more developed applications of sentiment analysis are particularly promising (Gayo-Avello 2015) and can achieve good results. The book is devoted precisely to this. First, we show how sentiment analysis can be performed through reliable and accurate techniques that might succeed in nowcasting and forecasting the electoral outcomes (i.e. the SASA methods presented in Table 1.1). Second, we distinguish between different types of analysis to see which design of the research is most suitable to decrease the error of the prediction.

Notes

[1] http://www.internetlivestats.com/internet-users/

[2] http://wearesocial.com/uk/special-reports/digital-social-mobile-worldwide-2015

[3] http://techcrunch.com/2012/11/09/can-social-media-influence-really-be-measured/

[4] https://www.article19.org/data/files/pdfs/publications/opinion-polls-paper.pdf

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., and Liu, B. (2012) ‘Online social networks flu trend tracker: A novel sensory approach to predict flu trends’, *Biomedical Engineering Systems and Technologies*: 357: 353–368.

Alvarez, R.M., and Hall, T.E. (2010) *Electronic Elections: The Perils and Promises of Digital Democracy*. Princeton, NJ: Princeton University Press.

Anduiza, E., Cantijoch, M., and Gallego, A. (2009) ‘Political participation and the internet: A field essay’, *Information, Communication & Society*, 12(6): 860–878.

Antweiler, W., and Frank, M.Z. (2004) ‘Is all that talk just noise? The information content of internet stock message boards’, *The Journal of Finance*, 59(3): 1259–1294.

Arzheimer, K., and Evans, J. (2014) ‘A new multinomial accuracy measure for polling bias’, *Political Analysis*, 22(1): 31–44.

Asur, S., and Huberman, B.A. (2010, August) ‘Predicting the future with social media’, *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 1: 492–499. IEEE.

Avery, J.M. (2009) ‘Videomalaise or virtuous circle? The influence of the news media on political trust’, *The International Journal of Press/Politics*, 14(4): 410–433. doi: 10.1177/1940161209336224

Bailard, C.S. (2012) ‘Testing the internet’s effect on democratic satisfaction: A multi-methodological, cross-national approach’, *Journal of Information Technology & Politics*, 9(2): 185–204.

Bakker, T.P., and De Vreese, C.H. (2011) ‘Good news for the future? Young people, internet use, and political participation’, *Communication Research*, 20: 1–20.

Barberá, P. (2015) ‘Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data’, *Political Analysis*, 23(1): 76–91.

Barberá, P., Popa, S.A., and Schmitt, H. (2015) ‘Prospects of ideological realignment (s) in the 2014 ep elections? Analyzing the common multidimensional political space for voters, parties, and legislators in Europe’, *MPSA Conference 2015*, Chicago, 16–19 April 2015.

Barclay, F.P., Pichandy, C., Venkat, A., and Sudhakaran, S. (2015) ‘India 2014: Facebook “like” as a predictor of election outcomes’, *Asian Journal of Political Science*, 23(2): 134–160. doi: 10.1080/02185377.2015.1020319

Barisione, M., and Ceron, A. (2016) ‘A digital movement of opinion? Criticizing austerity through social media’, In Barisione, M., and Michailidou, A. (eds.), *Social Media and European Politics: Rethinking Power and Legitimacy in the Digital Era*. Palgrave Macmillan.

Bastos, M.T., Mercea, D., and Charpentier, A. (2015) ‘Tents, tweets, and events: The interplay between ongoing protests and social media’, *Journal of Communication*, 65(2): 320–350.

Beauchamp, N. (2014) ‘Predicting and interpolating state-level polling using Twitter textual data’, Paper presented at the MPSA Annual National Conference, Chicago, March 2014.

Benkler, Y. (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press.

Bennett, W.L., and Segerberg, A. (2013) *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*. Cambridge: Cambridge University Press.

Bermingham, A., and Smeaton, A. (2011) 'On using Twitter to monitor political sentiment and predict election results', Paper presented at the Workshop on Sentiment Analysis Where AI Meets Psychology, Chiang Mai, Thailand, 13 November 2011.

Boireau, M. (2014) 'Determining political stances from Twitter timelines: The Belgian parliament case', *Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia*, St. Petersburg, Russia, 18–20 November 2014.

Bollen, J., Mao, H., and Zeng, X.J. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*, 2: 1–8.

Bond, R., and Messing, S. (2015) 'Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook', *American Political Science Review*, 109(1): 62–78.

Bothos, E., Apostolou, D., and Mentzas, G. (2010) 'Using social media to predict future events with agent-based markets', *IEEE Intelligent Systems*, 25(6): 50–58.

Boulianne, S. (2009) 'Does internet use effect engagement? A meta-analysis of research', *Political Communication*, 26(2): 193–211. doi: 10.1080/10584600902854363

Boutet, A., Kim, H., and Yoneki, E. (2013) 'What's in Twitter, I know what parties are popular and who you are supporting now!', *Social Network Analysis and Mining*, 3: 1379–1391.

Burnap, P., and Williams, M.L. (2015) 'Cyber hate speech on Twitter: An application of machine classification and statistical modelling for policy and decision making', *Policy & Internet*, 7(2): 223–242.

Butler, D. (2013) 'When google got flu wrong: Us outbreak foxes a leading web-based method for tracking seasonal flu', *Nature*, February 13. 494(7436): 155–156. doi:10.1038/494155a

Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M., and Riotta, G. (2014) 'A multi-level geographical study of Italian political elections from Twitter data', *PloS One*, 9(5): e95809.

Calderaro, A., and Kavada, A. (2013) 'Special issue on online collective action and policy change', *Policy & Internet*, 5(1): 1–6.

Callegaro, M., and Gasperoni, G. (2008) 'Accuracy of pre-election polls for the 2006 Italian parliamentary election: Too close to call', *International Journal of Public Opinion Research*, 20(2): 148–170.

Cameron, M.P., Barrett, P., and Stewardson, B. (2013) 'Can social media predict election results? Evidence from New Zealand', Working Paper in Economics 13/08, University of Waikato.

Campante, F., Durante, R., and Sobbrio, F. (2013) 'Politics 2.0: The multifaceted effect of broadband internet on political participation', *Working Paper*. Available at: http://www.iae.csic.es/investigatorsMaterial/a13174085038513.pdf

Ceron, A. (2015) 'Internet, news and political trust: The difference between social media and online media outlets', *Journal of Computer-Mediated Communication*, 20(5): 487–503.

Ceron, A. (2016a) 'Competing principals 2.0? The impact of Facebook in the 2013 selection of the Italian head of state', *Italian Political Science Review*, 46: 313–333. doi: http://dx.doi.org/10.1017/ipo.2016.14

Ceron, A. (2016b) ‘Intra-party politics in 140 characters: To what extent social media analysis provides information on intra-party dynamics? Three applications to the Italian case’, *Party Politics*. Advance online publication. doi: 10.1177/1354068816654325

Ceron, A., and Curini, L. (2016) ‘e-Campaigning in the 2014 European elections: The emphasis on valence issues in a two-dimensional multiparty system’, Advance online publication. *Party Politics*, doi: 10.1177/1354068816642807

Ceron, A., Curini, L., and Iacus, S.M. (2016) ‘First and second level agenda-setting in the Twitter-sphere: An application to the Italian political debate’, *Journal of Information Technology & Politics*, 13(2): 159–174.

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

Ceron, A., Curini, L., Iacus, S.M., and Ruggeri, A. (2015) ‘Here’s a paradox: Shutting down the Islamic State on Twitter might help it recruit’, *The Washington Post*, December 10. Available at: https://www.washingtonpost.com/news/monkeycage/wp/2015/12/10/heres-a-paradox-shutting-down-the-islamic-state-on-twitter-might-help-it-recruit/

Ceron, A., and d’Adda, G. (2015) ‘E-campaigning on Twitter: The effectiveness of distributive promises and negative campaign in the 2013 Italian election’, *New Media & Society*, 18(9), 1935–1955. doi: 10.1177/1461444815571915

Ceron, A., and Memoli, V. (2016) ‘Flames and debates: do social media affect satisfaction with democracy?’, *Social Indicators Research*, 126(1): 225–240.

Ceron, A., and Negri, F. (2016) ‘The “social side” of public policy: Monitoring online public opinion and its mobilization during the policy cycle’, *Policy & Internet*, 8(2): 131–147.

Chadwick, A. (2013) *The Hybrid Media System: Politics and Power*. New York, NY: Oxford University Press.

Choy, M., Cheong, M.L., Laik, M.N., and Shung, K.P. (2011) ‘A sentiment analysis of Singapore presidential election 2011 using Twitter data with census correction’, *arXiv preprint arXiv:1108.5520*. Available at: http://arxiv.org/abs/1108.5520

Choy, M., Cheong, M.L., Ma Nang, L., and Koo Ping, S. (2012) ‘US presidential election 2012 prediction using census corrected Twitter model’. Available at: http://arxiv.org/ftp/arxiv/papers/1211/1211.0938.pdf

Ciulla, F., Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N., and Vespignani, A. (2012) ‘Beating the news using social media: The case study of American Idol’, *EPJ Data Science*, 1(1): 1–11.

Conover, M., Ratkiewicz, J., Francisco, M.R., Gonçalves, B., Menczer, F., and Flammini, A. (2011) ‘Political polarization on Twitter’, *ICWSM*, 133: 89–96.

Conway, B.A., Kenski, K., and Wang, D. (2015) ‘The rise of Twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary’, *Journal of Computer Mediated Communication*, 20(4): 363–380.

Cook, S., Conrad, C., Fowlkes, A., and Mohebbi, M. (2011) ‘Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic’, *PLoS One*, 6(8): e23610.

Cosenza, V. (2013) ‘Elezioni 2013: i segnali in rete e il responso delle urne’, *Vincos Blog*. Available at: http://vincos.it/2013/02/26/elezioni-2013-i-segnali-in-rete-e-il-responso-delle-urne/

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

Cunha, E., Magno, G., Gonçalves, M.A., Cambraia, C.A., and Virgilio, A. (2014) ‘He votes or she votes? Female and male discursive strategies in Twitter political hashtags’, *PloS One*, 9(1): e87041.

Cunningham, J.A. (2012) 'Using Twitter to measure behavior patterns', *Epidemiology*, 23: 764–765.

Curini, L., Iacus, S., and Canova, L. (2015) 'Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case', *Social Indicators Research*, 121(2): 525–542.

Curini, L., and Splendore, S. (2016) 'The ideological proximity between citizens and journalists and its consequences: An application to the Italian case', Paper presented at Italian Political Science Society Annual Conference, Milan (Italy), 15–17 September 2016.

Curti, M., Iacus, S.M., Porro, G., and Siletti, E. (2015) 'Measuring social well being in the big data era: Asking or listening?', *arXiv preprint arXiv:1512.07271*.

Dekker, R., and Bekkers, V. (2015) 'The contingency of governments' responsiveness to the virtual public sphere: A systematic literature review and meta-synthesis', *Government Information Quarterly*, 32(4): 496–505.

de Voogd, L., Chelala, P., and Schwarzer, S. (2012) 'Do social media affect public discourses? A sentiment analysis of political tweets during the French presidential election campaign', Presented at the Annual Conference of the American Association of Public Opinion Research, Orlando, FL.

De Wilde P., Michailidou A., and Trenz H.-J. (2014) 'Converging on Euroscepticism: Online polity contestation during European parliament elections', *European Journal of Political Research* 53(4): 766–783.

Dimitrova, D.V., Shehata, A., Strömbäck, J., and Nord, L.R. (2014) 'The effects of digital media on political knowledge and participation in election campaigns: Evidence from panel data', *Communication Research*, 41: 95–118.

Druckman, J.N., Kifer, M.J., and Parkin, M. (2010) 'Timeless strategy meets new medium: Going negative on congressional campaign web sites, 2002–2006', *Political Communication*, 27(1): 88–103.

Durahim, A.O., and Coşkun, M. (2015) '#iamhappybecause: Gross national happiness through Twitter analysis and big data', *Technological Forecasting and Social Change*, 99: 92–105.

Ecker, A. (2015) 'Estimating policy positions using social network data: cross-validating position estimates of political parties and individual legislators in the Polish parliament', *Social Science Computer Review*. doi: 10.1177/0894439315602662

Evans, H.K., Cordova, V., and Sipole, S. (2014) 'Twitter style: An analysis of how house candidates used Twitter in their 2012 campaigns', *PS: Political Science & Politics*, 47(2): 454–462.

Falck, O., Gold, R., and Heblich, S. (2012) 'E-lections: Voting behavior and the internet', IZA DP No. 6545, Germany.

Farrell, H. (2012) 'The consequences of the internet for politics', *Annual Review of Political Science*, 15: 35–52.

Fink, C., Bos, N., Perrone, A., Liu, E., and Kopecky, J. (2013) 'Twitter, public opinion, and the 2011 Nigerian presidential election', *2013 International Conference on Social Computing (SocialCom)*, Washington, DC, 8–14 September 2013.

Fonseca, A. (2011) 'Modeling political opinion dynamics through social media and multi-agent simulation', First Doctoral Workshop for Complexity Sciences. Available at: http://idpcc.dcti.iscte.pt/docs/Papers_1st_Doctoral_Workshop_15–6–2011/AntonioFonseca.pdf

Franch, F. (2013) '(Wisdom of the crowds)^2: 2010 UK election prediction with social media', *Journal of Information Technology & Politics*, 10(1): 57–71.

Franco, A., Malhotra, N., and Simonovits, G. (2014) 'Publication bias in the social sciences: Unlocking the file drawer', *Science*, published online 28 August 2014. 345(6203), 1502–1505. doi: 10.1126/science.1255484

Freifeld, C.C., Mandl, K.D., Reis, B.Y., and Brownstein, J.S. (2008) 'HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports', *Journal of the American Medical Informatics Association*, 15(2): 150–157.

Gaurav, M., Kumar, A., Srivastava, A., and Miller, S. (2013) 'Leveraging candidate popularity on Twitter to predict election outcome', *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, Chicago, IL, 11 August 2013.

Gayo-Avello, D. (2011) 'Don't turn social media into another "literary digest" poll', *Communications of the ACM*, 54(10): 121–128.

Gayo-Avello, D. (2012) 'No, you cannot predict elections with Twitter', *IEEE Internet Computing*, 16(6): 91–94.

Gayo-Avello, D. (2013) 'A meta-analysis of state-of-the-art electoral prediction from Twitter data', *Social Science Computer Review*, 31(6): 649–679.

Gayo-Avello, D. (2015) 'Political opinion', In Mejova, Y., Weber, I., and Macy, M.W. (eds.), *Twitter: A Digital Socioscope*, 52–74. New York, NY: Cambridge University Press.

Ghiassi, M., Skinner, J., and Zimbra, D. (2013) 'Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network', *Expert Systems With Applications*, 40(16): 6266–6282.

Gibson, R.K., and McAllister, I. (2014) 'Normalising or equalising party competition? Assessing the impact of the web on election campaigning', *Political Studies*, 63(3): 529–547. doi: 10.1111/1467–9248.12107

Giglietto, F. (2012) 'If likes were votes: An empirical study on the 2011 Italian administrative elections', *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 4–7 June 2012.

Gilbert, E., and Karahalios, K. (2010, May) 'Widespread worry and the stock market', *Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, 23–26 May 2010.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2009) 'Detecting influenza epidemics using search engine query data', *Nature*, 457: 1012–1014.

Groshek, J., and Dimitrova, D. (2011) 'A cross-section of voter learning, campaign interest and intention to vote in the 2008 American election: Did Web 2.0 matter', *Communication Studies Journal*, 9: 355–375.

Gruhl, D., Guha, R., Kumar, R., Novak, J., and Tomkins, A. (2005) 'The predictive power of online chatter', *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, IL, 21–24 August 2005

Hanretty, C. (2011) 'MPs' Twitter activity reveals their ideology', Unpublished manuscript. Available at: http://chrishanretty.co.uk/blog/wp-content/uploads/2011/04/article1.pdf

Hansen, K.M., and Kosiara-Pedersen, K. (2014) 'Cyber-campaigning in Denmark: Application and effects of candidate campaigning', *Journal of Information Technology & Politics*, 11(2): 206–219.

Hansen, K.M., and Pedersen, R.T. (2008) 'Negative campaigning in a multiparty system', *Scandinavian Political Studies*, 31: 408–427.

Hanson, C.L., Burton, S.H., Giraud-Carrier, C., West, J.H., Barnes, M.D., and Hansen, B. (2013) 'Tweaking and tweeting: Exploring Twitter for nonmedical use of a psychostimulant drug (adderall) among college students', *Journal of Medical Internet Research*, 15(4): e62.

Helbing, D. (2015) *Thinking Ahead-Essays on Big Data, Digital Revolution, and Participatory Market Society*. Cham, Switzerland: Springer International Publishing.

Hermida, A., Lewis, S.C., and Zamith, R. (2014) 'Sourcing the Arab spring: A case study of Andy Carvin's sources on Twitter during the Tunisian and Egyptian revolutions', *Journal of Computermediated Communication*, 19(3): 479–499.

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. (2005) *Voice, Equality, and the Internet*. Princeton, NJ: Princeton University.

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

Hosch-Dayican, B., Aarts, K., Amrit, C., and Dassen, A. (2013) 'Issue Salience and Issue Ownership Online and Offline: Comparing Twitter and Survey Data', *APSA 2013 Annual Meeting Paper*.

Hosch-Dayican, B., Chintan, A., and Kees, A., (2016) 'How do online citizens persuade fellow voters? Using Twitter during the 2012 Dutch parliamentary election campaign', *Social Science Computer Review*, 34(2): 135–152.

Howard, P.N., and Hussain, M.M. (2011) 'The role of digital media', *Journal of Democracy*, 22(3): 35–48.

Huber, J., and Hauser, F. (2005) 'Systematic mispricing in experimental markets–evidence from political stock markets', *Proceedings of the International Conference on Finance*, Copenhagen, Denmark.

Huberman, B., and Asur, S. (2011). 'Predicting future outcomes', U.S. Patent Application No. 12/905,735.

Iacus, S.M. (2014) 'Big data or big fail? The good, the bad and the ugly and the missing role of statistics, Elec.', *Journal of Applied Statistical Analysis*, 5(11): 4–11.

Iacus, S.M., Porro, G., Salini, S., and Siletti, E. (2015) 'Social networks, happiness and health: From sentiment analysis to a multidimensional indicator of subjective well-being', *ArXiv*. Available at: http://arxiv.org/abs/1512.01569

Im, T., Cho, W., Porumbescu, G., and Park, J. (2014) 'Internet, trust in government, and citizen compliance', *Journal of Public Administration Research and Theory*, 24(3): 741–763.

Jacobs, K., and Spierings, N. (2016) 'Saturation or maturation? The diffusion of Twitter and its impact on voting in the Dutch general elections of 2010 and 2012', *Journal of Information, Technology and Politics*, 13: 1–21.

Jamal, A.A., Keohane, R.O., Romney, D., and Tingley, D. (2015) 'Anti-Americanism and anti-interventionism in Arabic Twitter discourses', *Perspectives on Politics*, 13(1): 55–73.

Jennings, M.K., and Zeitner, V. (2003) 'Internet use and civic engagement: A longitudinal analysis', *Public Opinion Quarterly*, 67(3): 311–334.

Jiang, Y. (2014) '"Reversed agenda-setting effects" in China case studies of Weibo trending topics and the effects on state-owned media in China', *Journal of International Communication*, 20(2): 168–183.

Jungherr, A. (2013, October) 'Tweets and votes, a special relationship: The 2009 federal election in Germany', *Proceedings of the 2nd Workshop on Politics, Elections and Data*, San Francisco, CA, 27 October–1 November 2013.

Jungherr, A. (2014) 'The logic of political coverage on Twitter: Temporal dynamics and content', *Journal of Communication*, 64(2): 239–259.

Jungherr, A., Jürgens, P., and Schoen, H. (2012) '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*, 30(2): 229–234.

Jungherr, A., Posegga, O., Schoen, H., and Jürgens, P. (2014) 'Attention online and electoral success: An uneasy relationship', Paper presented at Election Studies: Reviewing the Bundestagswahl 2013, Munich, 26–27 June 2014.

Jungherr, A., Schoen, H., Posegga, O., and Jurgens, P. (2016) 'Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support', Advance online publication. *Social Science Computer Review*: 1–21. doi: 10.1177/0894439316631043

Kaufhold, K., Valenzuela, S., and De Zúñiga, H.G. (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.

Kaye, B.K., and Johnson, T.J. (2002) 'Online and in the know: Uses and gratifications of the Web for political information', *Journal of Broadcasting & Electronic Media*, 46(1): 54–71.

Khatua, A., Khatua, A., Ghosh, K., and Chaki, N. (2015) Can *#Twitter_Trends* 'Predict Election Results? Evidence from 2014 Indian', Paper presented at the 48th Hawaii International Conference on System Sciences, Grand Hyatt, HI, 5–8 January 2015. doi: 10.1109/HICSS.2015.202

King, A., Orlando, F., and Sparks, D. (2011) 'Ideological extremity and primary success: A social network approach', Paper presented at the 2011 MPSA Conference, Chicago, USA, 31 March – 3 April.

King, G. (2014) 'Restructuring the social sciences: Reflections from Harvard's institute for quantitative social science', *Politics and Political Science*, 47(1): 165–172.

King, G., Pan, J., and Roberts, M.E. (2013) 'How censorship in China allows government criticism but silences collective expression', *American Political Science Review*, 107(2): 326–343.

Kobayashi, T., Ikeda, K.I., and Miyata, K. (2006) 'Social capital online, collective use of the internet and reciprocity as lubricants of democracy', *Information, Communication & Society*, 9: 582–611.

Koehler-Derrick, G., and Goldstein, J. (2011) 'Using Google insights to assess Egypt's Jasmine revolution', *CTC Sentinel*, 4(3): 4–8.

Ku, G., Kaid, L.L., and Pfau, M. (2003) 'The impact of web site campaigning on traditional news media and public information processing', *Journalism and Mass Communication Quarterly*, 80(3): 528–547.

Lampos, V. (2012) 'On voting intentions inference from Twitter content: A case study on UK 2010 general election'. Available at: arXiv:1204.0423v11

Lampos, V., and Cristianini, N. (2012) 'Nowcasting events from the social web with statistical learning', *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4): 72.

Lampos, V., Preotiuc-Pietro, D., and Cohn, T. (2013) 'Auser-centric model of voting intention from social media, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 4–9 August 2013.

Lansdall-Welfare, T., Lampos, V., and Cristianini, N. (2012) 'Nowcasting the mood of the nation', *Significance*, 9(4): 26–28.

Larsson, A.O. (2015) 'Pandering, protesting, engaging: Norwegian party leaders on Facebook during the 2013 "short campaign"', *Information, Communication & Society*, 18(4): 459–473.

Lee, J.K. (2007) 'The effect of the internet on homogeneity of the media agenda: A test of the fragmentation paper', *Journalism & Mass Communication Quarterly*, 84(4): 745–760.

Leetaru, K. (2011) 'Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space', *First Monday*, 16(9).

LeFebvre, R.K., and Armstrong, C. (2016) 'Grievance-based social movement mobilization in the #Ferguson Twitter storm" ', *New Media & Society*. doi: 10.1177/1461444816644697

Lev-On, A. (2011) 'Campaigning online: Use of the internet by parties, candidates and voters in national and local election campaigns in Israel', *Policy & Internet*, 3(1): 1–28.

Lewis, S.C. (2009) 'Citizen journalism: Motivations, methods, and momentum', In McCombs, M., Hinsley, A.W., Kaufhold, K., and Lewis, S.C. (eds.), *The Future of News: An Agenda of Perspectives*, 59–76. San Diego, CA: Cognella.

Lewis, S.C. (2012) 'The tension between professional control and open participation', *Information, Communication & Society*, 15(6): 836–866.

Lewis-Beck, M.S. (2005) 'Election forecasting: Principles and practice', *British Journal of Politics & International Relations*, 7: 145–164.

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

Liviu, L. (2011) 'Predicting product performance with social media', *Informatics in Education*, 15(2): 46–56.

Livne, A., Simmons, M.P., Adar, E., and Adamic, L.A. (2011) 'The party is over here: Structure and content in the 2010 election', *ICWSM*, 11: 17–21.

Lloyd, L., Kaulgud, P., and Skiena, S. (2006) 'Newspapers vs. blogs: Who gets the scoop?', In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 117–124.

MacWilliams, M.C. (2015) 'Forecasting congressional elections using Facebook data', *PS: Political Science & Politics*, 48(4): 579–583.

Malik, S. (2014) 'Support for Isis stronger in Arabic social media in Europe than in Syria, The Guardian, 28 November'. Available at: http://www.theguardian.com/world/2014/nov/28/support-isis-stronger-arabic-social-media-europe-us-than-syria

Martin, E., Traugott, M., and Kennedy, C. (2005) 'A review and proposal for a new measure of poll accuracy', *Public Opinion Quarterly*, 69(3): 342–369.

McCombs, M. (2005) 'A look at agenda setting: Past, present, and future', *Journalism Studies*, 6(4): 543–557.

McLaren, N., and Shanbhogue, R. (2011) 'Using internet search data as economic indicators', *Bank of England Quarterly Bulletin*, Q(2), 134–140.

McNeal, R., Hale, K., and Dotterweich, L. (2008) 'Citizen – government interaction and the internet: Expectations and accomplishments in contact, quality, and trust', *Journal of Information Technology & Politics*, 5(2): 213–229.

Meraz, S. (2009) 'Is there an elite hold? Traditional media to social media agenda setting influence in blog networks', *Journal of Computer-Mediated Communication*, 14: 682–707.

Meraz, S. (2011) 'The fight for "how to think": Traditional media, social networks, and issue interpretation', *Journalism*, 12(1): 107–127.

Meraz, S., and Papacharissi, Z. (2012) 'Networked gatekeeping and networked framing on #Egypt', *The International Journal of Press/Politics*, 18(2): 138–166.

Mercea, D. (2017) 'Building contention word-by-word: Social media usage in the European stop ACTA movement', In Barisione, M., and Michailidou, A. (eds.), *Social Media and European Politics: Rethinking Power and Legitimacy in the Digital Era*. London: Palgrave Macmillan.

Metaxas, P.T., Mustafaraj, E., and Gayo-Avello, D. (2011) ‘How (not) to predict elections’, *Proceedings of PASSAT/SocialCom 2011, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, IEEE Computer Society, Los Alamitos, CA, USA: 165–171.

Miner, L. (2012) ‘The unintended consequences of internet diffusion: Evidence from Malaysia’, *Journal of Public Economics*, 132, 66-78.

Morozov, E. (2011) *The Net Delusion: The Dark Side of Internet Freedom*. New York: Public Affairs.

Murphy, J.J., Kim, A., Hansen, H.M., Richards, A.K., Augustine, C.B., Kroutil, L.A., and Sage, A.J. (2011). ‘Twitter feeds and Google search query surveillance: Can they supplement survey data collection?’, *Proceedings of the Association for Survey Computing Sixth International Conference*, Bristol, 22–23 September 2011.

Mustafaraj, E., Finn, S., Whitlock, C., and Metaxas, P.T. (2011) ‘Vocal minority versus silent majority: Discovering the opinions of the long tail’, *Proceedings of PASSAT/SocialCom 2011, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, IEEE Computer Society, Los Alamitos, CA: 103–110.

Neuman, R.W., Guggenheim, L., Mo Jang, S., and Young Bae, S. (2014) ‘The dynamics of public attention: Agenda-setting theory meets big data’, *Journal of Communication*, 64(2): 194–214.

Nisbet, E.C., Stoycheff, E., and Pearce, K.E. (2012) ‘Internet use and democratic demands: A multinational, multilevel model of internet use and citizen attitudes about democracy’, *Journal of Communication*, 62: 249–265.

Nooralahzadeh, F., Arunachalam, F., and Chiru, C. (2013) ‘Presidential elections on Twitter – An analysis of how the US and French election were reflected in tweets’, Paper presented at the 19th International Conference on Control Systems and Computer Science, Bucharest, Romania, 29–31 May 2013. doi: 10.1109/CSCS.2013.72

Norris, P. (2011) *Democratic Deficit: Critical Citizens Revisited*. Cambridge, UK: Cambridge University Press.

O’Connor, B., Balasubramanyan, R., Routledge, B.R., and Smith, N.A. (2010) ‘From tweets to polls: Linking text sentiment to public opinion time series’, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, 23–26 May.

Oliveira, D.J.S., de Souza Bermejo, P.E., and dos Santos, P.A. (2015) ‘Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls’, *Journal of Information Technology & Politics*. doi: 10.1080/19331681.2016.1214094

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., and de Fatima Oliveira, M. (2012) ‘Affective news and networked publics: The rhythms of news storytelling on #Egypt’, *Journal of Communication*, 62(2): 266–282.

Parmelee, J.H. (2013) ‘The agenda-building function of political tweets’, *New Media & Society*, 16(3): 434–450.

Parmelee, J.H., and Bichard, S.L. (2011) *Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public*. Lanham, MD: Lexington Books.

Pew Research Center. (2012) ‘Social media and voting’. Available at: http://www.pewinternet.org/files/old-media//Files/Reports/2012/PIP_TheSocialVote_PDF.pdf

Preis, T., Moat, H.S., and Stanley, H.E. (2013) ‘Quantifying trading behavior in financial markets using Google trends. Scientific reports 3 1684’. Available at: http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html

Puglisi, R., and Snyder, J.M. (2016) ‘Empirical studies of media bias’, In Anderson, S., Waldfogel, J., and Stromberg, D. (eds.), *Handbook of Media Economics*, 647–667. Amsterdam: Elsevier.

Quintelier, E., and Vissers, S. (2008) ‘The effect of internet use on political participation’, *Social Science Computer Review*, 26(4): 411–427.

Radinsky, K., and Horvitz, E. (2013) ‘Mining the web to predict future events’, *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, Rome, Italy, 4–8 February 2013.

Reilly, S., Richey, S., and Taylor, J.B. (2012) ‘Using Google search data for state politics research: An empirical validity test using roll-off data’, *State Politics & Policy Quarterly*, 12(2): 146–159.

Reuter, O.J., and Szakonyi, D. (2015) ‘Online social media and political awareness in authoritarian regimes’, *British Journal of Political Science*, 45(1): 29–51.

Roberts, M., Wanta, W., and Dzwo, T.-H. (2002) ‘Agenda setting and issue salience online’, *Communication Research*, 29(1): 452–465.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2013) ‘Tweet analysis for real-time event detection and earthquake reporting system development’, *IEEE Transactions on Knowledge and Data Engineering*, 25(4): 919–931.

Sanders, E., and van den Bosch, A. (2013) ‘Relating political party mentions on Twitter with polls and election results’. Available at: ceur-ws.org/Vol-986/paper_9.pdf

Sayre, B., Bode, L., Shah, D., Wilcox, D., and Shah, C. (2010) ‘Agenda setting in a digital age’, *Policy & Internet*, 2(2): 7–32.

Scharkow, M., and Vogelgesang, J. (2011) ‘Measuring the public agenda using search engine queries’, *International Journal of Public Opinion Research*, 23(1): 104–113.

Scheufele, D.A., and Nisbet, M.C. (2002) ‘Being a citizen online new opportunities and dead ends’, *The Harvard International Journal of Press/Politics*, 7(3): 55–75.

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Lucas, R.E., Agrawal, M., Park, G.J., Lakshmikanth, S.K., Jha, S., Seligman, M.E., and Ungar, L.H. (2013) ‘Characterizing geographic variation in well-being using tweets’, *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, 8–11 July 2013.

Schweitzer, E.J. (2008) ‘Innovation or normalization in e-campaigning? A longitudinal content and structural analysis of German party websites in the 2002 and 2005 national elections’, *European Journal of Communication*, 23: 449–470.

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

Senkyo kakaricho (2009) ‘Summary of reviews predicted that election results’. Available at: http://senkyo.kakaricho.jp/2009/report1.html

Shi, L., Agarwal, N., Agrawal, A., Spoelstra, G., and Spolestra, J. (2012) ‘Predicting US primary elections with Twitter’, Unpublished manuscript. Available at: http://snap.stanford.edu/social2012/papers/shi.pdf

Shirky, C. (2011) ‘The political power of social media’, *Foreign Affairs*, 90(1): 28–41.

Signorini, A., Segre, A.M., and Polgreen, P.M. (2011) ‘The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic’, *PloS One*, 6(5): e19467.

Skaperdas, S., and Grofman, B. (1995) ‘Modeling negative campaigning’, *American Political Science Review*, 89(1): 49–61.

Stephens, M. (2013) ‘The geography of hate’. Available at: http://www.floatingsheep.org/2013/05/hatemap.html

Stoycheff, E., and Nisbet, E.C. (2014) ‘What’s the bandwidth for democracy? Deconstructing internet penetration and citizen attitudes about governance’, *Political Communication*, 31(4): 628–646.

Strandberg, K. (2013) ‘A social media revolution or just a case of history repeating itself? The use of social media in the 2011 Finnish parliamentary elections’, *New Media & Society*, 15(8): 1329–1347.

Sudulich, M.L., and Wall, M. (2010) ‘Every little helps: Cyber campaigning in the 2007 Irish general election’, *Journal of Information Technology and Politics*, 7(4): 340–355.

Sunstein, C.R. (2001) *Republic.com*. Princeton: Princeton University Press.

Surowiecki, J. (2004) *The Wisdom of Crowds*. New York: Doubleday.

Swigger, N. (2013) ‘The online citizen: Is social media changing citizens’ beliefs about democratic values?’, *Political Behavior*, 35(3): 589–603.

Sylwester, K., and Purver, M. (2015) ‘Twitter language use reflects psychological differences between democrats and republicans’, *PloS One*, 10(9): e0137422.

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, 23 April 2012.

Tolbert, C.J., and McNeal, R.S. (2003) ‘Unraveling the effects of the internet on political participation?’, *Political Research Quarterly*, 56(2): 175–185.

Tolbert, C.J., and Mossberger, K. (2006) ‘The effects of e-government on trust and confidence in government’, *Public Administration Review*, 66(3): 354–369.

Tsakalidis, A., Papadopoulos, S., Cristea, A., and Kompatsiaris, Y. (2015) ‘Predicting elections for multiple countries using Twitter and polls’, *IEEE Intelligent Systems*, 30: 10–17. doi: 10.1109/MIS.2015.17

Tufekci, Z., and Wilson, C. (2012) ‘Social media and the decision to participate in political protest: Observations from Tahrir square’, *Journal of Communication*, 62(2): 363–379.

Tumasjan, A., Sprenger, T.O., Philipp, G.S., and Welpe, I.M. (2010) ‘Predicting elections with Twitter: What 140 characters reveal about political sentiment’, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, 23–26 May 2010.

Unankard, S., Li, X., Sharaf, M., Zhong, J., and Li, X. (2014) ‘Predicting elections from social networks based on sub-event detection and sentiment analysis’, *Web information systems engineering – WISE 2014*: 1–16. New York: Springer International Publishing.

UzZaman, N., Blanco, R., and Matthews, M. (2012) ‘TwitterPaul: Extracting and aggregating Twitter predictions’, *Artificial Intelligence; Physics and Society*. Available at: http://arxiv.org/abs/1211.6496.

Vaccari, C., Valeriani, A., and Barberá, P. (2013) ‘Social media and political communication: A survey of Twitter users during the 2013 Italian general election’, *Italian Political Science Review*, 43(3): 381–409.

Valdivia, A., Lopez-Alcalde, J., Vicente, M., Pichiule, M., Ruiz, M., and Ordobas, M. (2010) ‘Monitoring influenza activity in Europe with Google flu trends: Comparison with the findings of sentinel physician networks-results for 2009–10’, *Eurosurveillance*, 15(29): 2–7.

Valenzuela, S., Namsu, P., and Kerk, F.K. (2009) ‘Is there social capital in a social network site?: Facebook use and college students’ life satisfaction, trust, and participation’, *Journal of Computer-Mediated Communication*, 14(4): 875–901.

Vargo, C.J. (2011) ‘Twitter as public salience: An agenda-setting analysis’, Paper presented at the annual conference of AEJMC, St. Louis, MO.

Vargo, C.J., Basilaia, E., and Shaw, D.L. (2015) ‘Event versus issue: Twitter reflections of major news, a case study’, *Communication and Information Technologies Annual – Studies in Media and Communications*, 9: 15–239.

Vergeer, M., and Hermans, L. (2013) ‘Campaigning on Twitter: Micro-blogging and online social networking as campaign tools in the 2010 general elections in the Netherlands’, *Journal of Computer-Mediated Communication*, 18(4): 399–419.

Vergeer, M., Hermans, L., and Sams, S. (2013) ‘Online social networks and micro-blogging in political campaigning: The exploration of a new campaign tool and a new campaign style’, *Party Politics*, 19(3): 477–501.

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

Wallsten, K. (2007) ‘Agenda setting and the blogosphere: An analysis of the relationship between mainstream media and political blogs’, *Review of Policy Research*, 24(6): 567–587.

Wallsten, K. (2010) ‘“Yes we can”: How online viewership, blog discussion, campaign statements, and mainstream media coverage produced a viral video phenomenon’, *Journal of Information Technology & Politics*, 7(2–3): 163–181.

Washington, A.L., Parra, F., Thatcher, J.B., LePrevost, K., and Morar, D. (2013) ‘What is the correlation between Twitter, polls and the popular vote in the 2012 presidential election?’, *APSA 2013 Annual Meeting Paper*, Washington, DC.

Weeks, B., and Southwell, B. (2010) ‘The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and presidential politics’, *Mass Communication and Society*, 13(4): 341–360.

Weinberger, S. (2011) ‘Spies to use Twitter as crystal ball’, *Nature*, 478(7369): 301.

Welch, E.W., Hinnant, C.C., and Moon, M.J. (2005) ‘Linking citizen satisfaction with e-government and trust in government’, *Journal of Public Administration Research and Theory*, 15(3): 371–391.

Whyte, Christopher E. (2015) ‘Thinking inside the (black) box: Agenda setting, information seeking, and the marketplace of ideas in the 2012 presidential election’, *New Media & Society*, 18(8): 1680–1697. doi: 10.1177/1461444814567985

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*, Boston, MA, 28–31 August 2008.

Woodly, D. (2008) ‘New competencies in democratic communication? Blogs, agenda setting and political participation’, *Public Choice*, 134(1–2): 109–123.

Wu, M. (2012) ‘Big data, big prediction? – Looking through the predictive window into the future’. Available at: http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Big-Prediction-Looking-through-the-Predictive-Window/ba-p/41068

Yasseri, T., and Bright, J. (2015) ‘Predicting elections from online information flows: Towards theoretically informed models’, *ArXiv*. Available at: http://arxiv.org/abs/1505.01818

Yu, Y., Duan, W., and Cao, Q. (2013) ‘The impact of social and conventional media on firm equity value: A sentiment analysis approach’, *Decision Support Systems*, 55(4): 919–926.

Zeitzoff, T., Kelly, J., and Lotan, G. (2015) 'Using social media to measure foreign policy dynamics: An empirical analysis of the Iranian–Israeli confrontation (2012–13)', *Journal of Peace Research*, 52(3): 368–383. doi: 0022343314558700

Zhang, X., Fuehres, H., and Gloor, P.A. (2011) 'Predicting stock market indicators through Twitter "I hope it is not as bad as I fear"', *Procedia-Social and Behavioral Sciences*, 26: 55–62.

Zhang, X., Fuehres, H., and Gloor, P.A. (2012) 'Predicting asset value through Twitter buzz', In Altman, J., Baumöl, U., and Krämer, B.J. (eds.), *Advances in Collective Intelligence 2011*, 22–34. New York: Springer.

2 From noise to signal in sentiment and opinion analysis

From text analysis to opinion discovery

"*What do people think of . . . ?*" is the central question of decision makers, such as politicians, administrators, company managers or researchers in the social science. As seen in several examples in Chapter 1, the Internet and the vastitude of content in social media in particular are inexhaustible sources of data containing valuable information. But even before the outburst of the World Wide Web and social media, linguists, along with statisticians and computer scientists, have adapted old techniques and developed new ones in order to extract the *sentiment* and the *opinions* from digital texts. As in every field of research, each technique has its pros and cons and, in fact, there is no "best technique" or a universal one, although it is still possible to identify reasonable and efficient techniques.

Quantitative and qualitative analysis of texts

One of the most common mistakes of beginners in sentiment analysis (the analysis of *positive* or *negative* content in a text) or opinion analysis (*why* positive/negative?) in general is to rely on a brute force approach, letting computers automatically extract metrics, like counting how many times a certain word or hashtag appears, the number of *followers* of a Twitter account, the number of likes of a post on Facebook and so forth (see the discussion in Chapter 1). Those metrics are, in fact, useful information to a certain extent, for instance they could be useful to study the structure of a given personal/group network of relationships and its strenghts, but they do not produce any useful information in terms of *sentiment* or *opinion*. To go back to one previously discussed example, if we were simply focused on counting the number of *followers* of the two contenders for the White House in the presidential election of 2012, Obama with about 17 million followers, would have won 94 to 6 (with a gap of 88 percentage points) against the candidate Romney who had less than 1 million followers at the beginning of the campaign. The final difference between the two was less than 4% (see Chapter 3). So a mere quantitative analysis is not effective and may even be misleading, as in the previous case or the one of the Italian primary elections of the centre-left in November 2012. In that primary election, Renzi recorded 40,000 mentions on Twitter, 81%

of the total mentions, against about 15,000 of his opponent Bersani. According to the metric-logic Renzi should have won over Bersani by 10% in the first ballot and more than 20% in the second ballot. In actuality, Bersani won by 20% (see Chapter 3). Therefore, we can not attribute to social media or *sentiment analysis* the ineffectiveness of these predictions, but rather to the specific method chosen to analyze the data. Hence, a type of analysis that is able to combine the accuracy of quantitative estimation over Big Data with a more qualitative approach, through which we can dig deeper in the data, becomes necessary. It has to be recognized first that, so far, the best "machine" able to capture the subtle shades of natural language is the human brain. Moreover, the meaning of words is often interpretable only within a specific context, particularly in the political realm where the very definitions of terms are points of debate. If so, it is now time to reverse the Big Data paradigm from computer-powered analysis to human-driven data science.

The fundamental principles of text analysis

What do we mean by *sentiment analysis* and *opinion analysis* in this book? The first term is related to the estimation of the intensity (positive/negative) of a feeling (*sentiment*). The second is about the motivations behind this feeling (why positive? why negative?) as well as the identification of the topics linked to any discussion online. The sentiment analysis is closely related to the concept of *opinion mining*, a term introduced by Dave et al. (2003) to indicate a technique capable of conducting research on a given set of keywords and to identify attributes for each term (positive, neutral, negative), and once aggregating the distributions of these terms, it would be possible to extract the opinion associated to each keyword. Almost all the subsequent works have focused on producing essentially *sentiment* classification linked to individual keywords (Pang and Lee 2004; Pang et al. 2002). The term *sentiment analysis* has also replaced the term *text mining* which actually refers to textual analysis in a broad sense (Liu 2006). As already mentioned, by *opinion analysis* we mean the ability to extract the reasons behind a positive or negative *sentiment*.

The technique iSA (integrated Sentiment Analysis), derived from the fundamental work of Hopkins and King (2010) is, as we shall see, a technique of estimation of integrated opinions and feelings built on valid statistics grounds and not relying on computer science alone (Ceron et al. 2016). But just as no body escapes Newton's laws, no technique can escape the following fundamental principles of text analysis (Grimmer and Stewart 2013).

Principle 1: every quantitative linguistic model is wrong, but some can be useful

The mental process that leads to the production of a text, without exception, is purely a mystery. Even for the finest linguists or psychologists. Every single sentence can drastically change its meaning due to the inclusion or exclusion of very small bits of language. The following text is an authentic review of a film on the

Internet Movie Database (IMDb) portal: "*This film has good premises. Looks like it has a nice plot, an exceptional cast, first class actors and Stallone gives his best. But it sucks.*" The last three words (and, especially, the term *but*) completely change the meaning of that statement in the previous sentence. In fact, although there is a predominance of positive terms (five positive against one negative term), that three-word sentence completely reverses the semantic content of the whole text. Most of the time, it is the use of the so-called *functional words* (articles, nouns, prepositions, etc.) that characterize a sentence, although these functional terms alone do not mean anything and are usually discarded in automated sentiment extraction methods (Pennebaker et al. 2015). This is typical in Twitter data, where the use of a hashtag, somehow ironically, at the end of a comment can completely change the interpretation of the tweet. This is also the motivation why, in many situations, the "bag of words" approach (i.e. one that considers a text as represented by a vector of word counts or occurrences, without any reference to its grammatical structure) seems more useful than the structured one. Oftentimes it is the positioning or the absence of punctuation that can distort the meaning of a text, but in typical text analysis, punctuation is usually removed. Think about the famous prophecy attributed to Sibyl Latin: "*Ibis redibis numquam peribis in bello*", which can be translated as "*will go, will come back, will not die in war*", but also the opposite way, "*will go, will not come back, will die in war*". Word jokes, methaporic or ironic sentences like "*there is no favorable wind for the mariner who doesn't know where to go*" (Seneca), double meaning terms, and so forth make natural language interpretation great fun for humans and a painful job for computers.

Principle 2: quantitative methods help, but cannot replace, humans

For the reasons mentioned earlier, the automatic methods can make only some operations over texts faster and allow them to scale up to large corpora. They can be considered a tool that enhances or aids human capabilities (such as a telescope or a lever), but they are certainly not a tool to replace the human brain.

Principle 3: there exists no best or ideal technique of text analysis

Each technique is designed with very specific purposes and is based on assumptions defined a priori. Further, in the case of text analysis, there are additional constraints such as the *language* itself (English, Japanese, Spanish, etc.), the *topic* of discussion (politics, economy, sports), the *historical* period (the same words can be *hot*[1] or *cold*, depending on the historical period), the *age* and *gender* of the writer and the *nature* of the interlocutors (imagine a text in which two students, or a student and a teacher or two teachers discuss exams), and so forth. Moreover, there are also techniques designed for *individual* classification of texts and other designed for *aggregated* classification. For individual classification algorithms, the aim is to attribute a semantic category (or an author, a topic, etc.) to each unread text. Whereas in aggregated classification, the object of interest is to study the aggregate distribution of semantic categories (or opinions, topics, etc.) over the population of texts. In short, rather than search for the proverbial needle in the haystack (individual classification), we try to figure out the form the haystack takes (aggregated classification). In social science, *opinion* and *sentiment analysis* are intrinsically linked to the concept of aggregated classification, especially in forecasting tasks. In fact, it is usually more important to know who wins the elections rather than who each elector voted for. One could argue that the distinction between individual and aggregate classification is artificial, since the individual classification, after all, can also be used to produce the aggregate distribution. Unfortunately, the ex post aggregation is highly dangerous, if not detrimental, as we discuss later.

Principle 4: validation of the analysis

Every new method, as well as every model, must be validated by the data themselves. The supervised methods, that is those for which the semantic categories are known a priori or are identified by manual coding on a subset of texts called the *training set*, can easily be validated in every analysis by cross-validation, especially if we consider the individual classification techniques. This validation can be performed by checking the semantic classification generated by the method and the objective (or hand coded) semantic meaning of each texts. For unsupervised methods, where the semantic categories are identified *a posteriori* through the observation of recurrences within groups of texts classified as homogeneous or the assignment is made by cross-checking dictionaries of terms or catalogs, the validation is a particularly difficult task, if not impossible on a large scale. In such circumstances, the analysis may require the construction of controlled experiments, such as entering text for which you know the semantic content but of which the algorithm ignores the classification, and verify that the method assigns the document to the group that it is assumed to be correct.

How to make a text digestible to statistical models: the stemming

So far, we pointed out two main features that characterize the techniques of text: the *type of output* (or the target), that is individual or aggregated distribution estimation, and the *type of learning approach*, that is *supervised* or *unsupervised*. Among those methods which try to achieve individual classification there are also the subgroup of scoring methods, that is those methods which try to align the text on a fictitious line (left–right, etc.). Table 2.1 provides an initial summary of this classification, which will be useful at a later stage.

Before going into the details of every single techinque, let's begin with the *preprocessing* phase in order to see how a text is transformed into digital data so that an algorithm can then treat it.

We have already discussed the complexity of the language, but fortunately not all of this complexity is necessary for textual analysis. The initial, but crucial, process is the reduction of the texts into quantitative data so that they can be analyzed by a proper statistical model. A text contains many words or auxiliary

Table 2.1 What characterizes text analysis techniques

		Type of learning approach	
		unsupervised	*supervised*
Type of output	*individual*	individual analysis paired with ontological dictionaries	individual analysis paired with human coding or pre-tagged data
opinion/ sentiment estimation	*aggregated*	aggregated distribution estimation paired with ontological dictionaries	aggregated distribution estimation paired with human coding or pre-tagged data

symbols that can be filtered through the preliminary analysis. In general, a *text* or a *document* is part of a set of texts called the *corpus* and a collection of *corpus* is called *corpora*. There are algorithms that are more efficient on short texts and others that work best with longer texts but, regardless of the length, all methods include a similar process of text transformation into data matrix. One of the first procedures is the so-called *preprocessing* phase of the texts, which deletes the information about the order in which the words appear in the text (Jurafsky and Martin 2009). After the initial premises, this may seem very strange and unwanted but, in fact, the second principle discussed earlier still applies. According to this principle, an excessive super-structure of the linguistic model inevitably leads to results that will mimic the assumptions of the model and this will, in turn, make the whole text analysis process merely a tautology. This stemming phase which does not take into account the order of words in a sentence is referred to as *bag of words*. Furthermore, it is usual practice to reduce the text to a smaller set of terms called *stem*, which consists in the atomic data to analyze. *Stem* means a single word (*unigrams*) or, if one wants to give some importance to the ordering of words, a pair of words (*bigrams*) (i.e. the bigram "white house" in "The *White House* in Washington DC" will be treated differently from the couple of stems "I own a *white* and *small* house by the sea") or triads of words (*trigrams*), and so forth. Generally speaking, considering stems with three or more words does not provide an addition of information and does not increase the quality of the classification and, in fact, the most used stemming procedures are limited to unigrams. Stems are not necessarily whole words, but it is instead preferred to reduce them to its fundamental common roots, for example *family*, *families*, and *familiar* can be unified into the stem "*famil*". All conjunctions, punctuation, articles, prepositions, suffixes and prefixes, verb endings and so forth as well as words that appear *too frequently* within a *corpus* (for instance in 95% of texts) or *too rarely* (less than 5% of texts) may also be removed. This *stemming* step can be designed and optimized specifically for each language and robust and efficient tools exist nowadays.

Let us proceed with a fictitious example to explain the process. Let us suppose to have a *corpus* of texts regarding the *sentiment* about the possibility of reintroducing nuclear power plants in Italy. Suppose that the first three texts in the corpus are as follows:

- Text #1: "The **nuclear** is worthwhile because it's **cheap**"
- Text #2: "The **nuclear** energy produces **waste**"
- Text #3: "The **nuclear** power **scares** me because of **radiation**, the **waste** and does not reduce **pollution**"

Let's suppose that the *stemming* stage has kept only the words marked in bold. Now all the texts within the *corpus* are transformed into the rows of a matrix (document-term matrix) where each row represents a text and the columns represent the stems. If a word/stem appears in a given text, the cell of the document-term matrix will have a 1 in the corresponding row and 0 otherwise. It is also possible to use the counts instead of a binary coding, but experience shows that this does not significantly increase the reliability of the statistical analysis which follows the *stemming* phase.

Let's go back to the previous example: imagine that the stem s_1 is "*nuclear*", s_2 = "*scares*", s_3 = "*radiation*", s_4 = "*pollution*", s_5 = "*waste*", s_6 = "*cheap*", and so forth. Consider now the first text, $i = 1$, "the *nuclear* is worthwhile because it's *cheap*". We can associate to the vector of stem $S_1 = (s_1, s_2, s_3, s_4, s_5, s_6, \ldots) = (1, 0, 0, 0, 0, 1, \ldots)$; for text $i = 2$ the vector of stem is $S_2 = (1, 0, 0, 0, 1, 0, \ldots)$ and, in general, to the generic *i-th* text in the *corpus* a vector of *stem* will be denoted by S_i. Every text belongs to one semantic category D_k, $k = 1, \ldots, K$, where K is the total number of semantic categories for these data. For instance, to simplify the exposition we can set $K = 2$ and identify by D_1 = "*in favor*" of the return of nuclear power in Italy and D_2 = "*against*". The document-term matrix in Table 2.2 shows the *stem* vectors S_i, associated with the hand-coded *tag* of each text, figuring that only some of these texts have actually been already classified (in the example, the text #2 contains "NA", or Not Available, in column D_i as it was not manually tagged but the *stemming* can still be performed).

Table 2.2 Example of document-term matrix

Post	D_i	*Stem* s_1 *nuclear*	*Stem* s_2 *scares*	*Stem* s_3 *radiation*	*Stem* s_4 *pollution*	*Stem* s_5 *waste*	*Stem* s_6 *cheap*	...
text #1	in favor	1	0	0	0	0	1	...
text #2	NA	1	0	0	0	1	0	...
text #3	against	1	1	1	1	1	0	...
text #4	against	1	1	1	1	1	0	...
text #5	in favor	1	0	1	1	1	0	...
...	...	...	...	...	...	...	...	...
text #10000	in favor	1	0	1	0	0	1	...
...	...	...	...	...	...	...	..‥	...

Thinking about the number of words every language is made of, one might think that the document-term matrix might possess a huge number of rows in any given analysis. For instance, the Italian language consists of a number of terms that oscillates between 215,000 and 270,000, depending on the dictionary adopted. The *Oxford English Dictionary* classifies more than 650,000 words. What turns out to be true is that, after stemming, the typical length of the stem vector S_i is no more than 300 or 500 and often much less. The main dimension that increases the computational challenge is the number of rows of the matrix, that is the number of texts to be analyzed. This number is usually in the order of millions in social media analysis.

At this stage, we can assume that the document-term matrix is available and we are ready to review more in details several families of textual analysis methods. For this brief review, we refer to Figure 2.1, largely taken from Grimmer and Stewart (2013). This figure represents a rough ontology which extends the classification presented in Table 2.2.

Scoring techniques

The idea of *scoring* consists of ordering texts along a spectrum of continuous opinions rather than *classifying* them according to a discrete and finite set of categories. Scoring techniques refer to a more general framework called *Item Response Theory*, or IRT (de Boeck and Wilson 2004), which originated in psychometry and psychology. This theory assumes the existence of latent dimensions, which are the fictitious axes on which text lies. A typical application of these scoring techniques in political analysis is that of ordering electoral speeches or those documents written by political actors (Ceron 2015; Curini et al. 2013), for example the *scoring* along the left–right axis. These techniques can be implemented in both *supervised* or *unsupervised* approach. Among the unsupervised techniques, we can mention *Wordfish* (Slapin and Proksch 2008). This technique produces an ordering based on the frequency of terms contained in the texts but does not provide any clue on the meaning of the underlying axis. Indeed, once the ordering is available, the outcome should be analyzed further by looking at the texts which have been classified at extremes of the axis and infer from a qualitative analysis on them, possible polarities like positive/negative, left/right, materialism/post-materialism and so forth. More precisely, for each text i, the frequency y_{ij} of a word j is assumed to have a Poisson distribution with constant rate λ_{ij}, for example $y_{ij} \approx \text{Poisson}(\lambda_{ij})$, where $\lambda_{ij} = \exp(\alpha_i + \psi_j + \beta_j * \omega_i)$, and α_i is the fixed effect for document i, ψ_j is fixed effect for word j and β_j is a specific weight/importance of word j in discriminating documents and, finally, ω_i is the position of document i along the latent dimension. The estimation algorithm makes use of the EM (expectation–maximization) method, where the initial value of ψ_j is the log of the mean count for each word j, the initial value for α_i is relative log-ratio of the mean word count of each document i with respect to the first document in the corpus and the initial values of β_j and ω_i are obtained through SVD (singular value decomposition). Once the algorithm has been initialized, the estimation step proceeds by first estimating

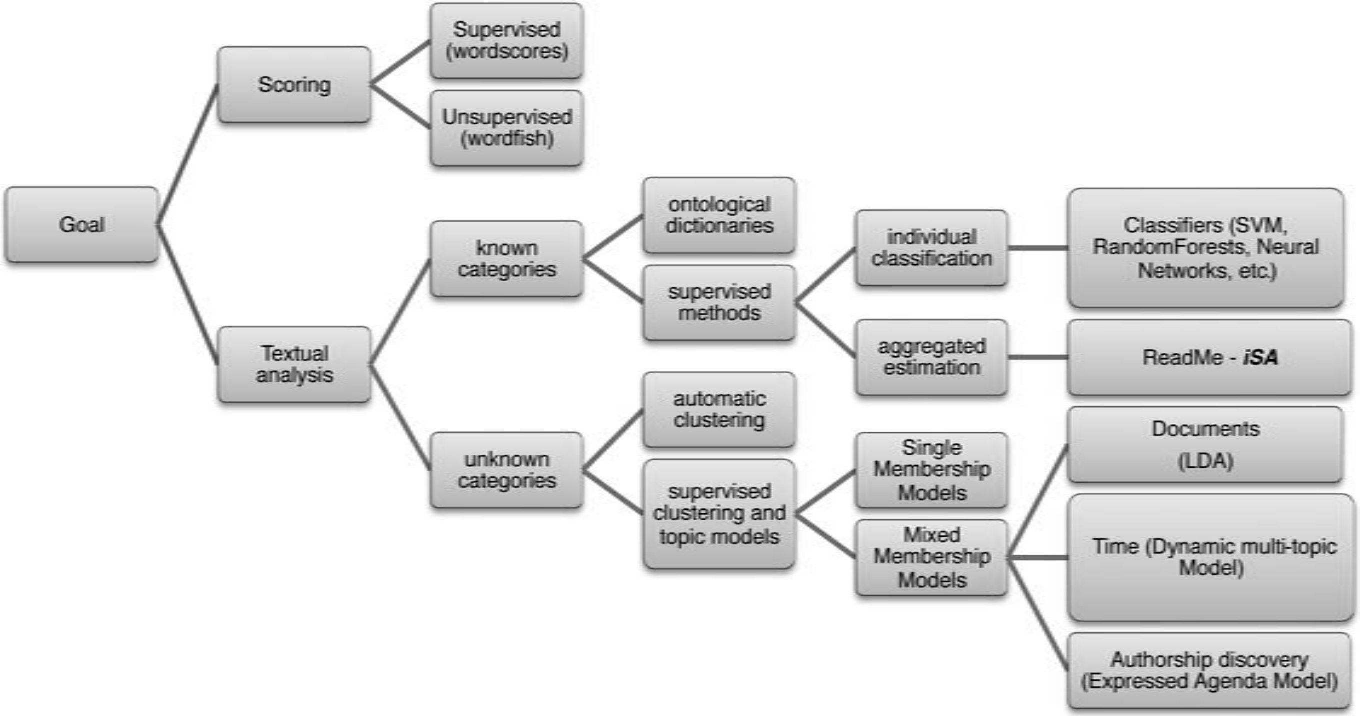

Figure 2.1 Needs and solutions in textual analysis

the couple (α_i, ω_i) conditionally on the previous values of (ψ_j, β_j) by maximum likelihood and then, by using the estimates of (α_i, ω_i) at this step, the new values of (ψ_j, β_j) are obtained using a Gaussian $N(0, \sigma^2)$ prior for β_j. Finally, a global likelihood is evaluated and the procedure iterated till convergence. This approach has several limitations though. For example if texts are too small, the method is unstable and convergence is not granted. Moreover, if the value of σ is too small, each single word can split two documents apart. Finally, it is computationally intensive and the results do not allow for an intuitive interpretation of the latent variable, as the method is completely unsupervised. Despite these limitations, the method is still valuable, as it is not dictionary based and hence works with texts written in any language.

One of the most famous supervised scoring algorithms is called *Wordscores* (Laver et al. 2003). This method assumes that there exists a source of already classified texts along one or more dimensions and, therefore, produces a scoring along these predefined axes. So, while *Wordfish* is more an explorative data analysis tool, Wordscores is closer to a machine learning algorithm. We do not go into the details of this method, but we refer to the original papers (Benoit and Laver 2003; Laver et al. 2003).

Pros and cons of human and automatic tagging

We remind readers that by *supervised* algorithm we mean here a procedure which attributes a semantic meaning to each text in a corpus given a precoded set of texts. This step is also called *tagging*, and tagging may occur through human coding or automatic coding based on, for example ontological dictionaries. An ontological dictionary is nothing but a collection of terms categorized according to some semantic macro-areas. Ontological dictionaries *per se* are not that useful. Indeed, tagging through ontological dictionaries also requires a set of rules which explain how a combination of words or stems appearing a text can be associated to a semantic meaning. Think about the phrase "*what a nice rip-off*". The term *nice* (positive adjective) and *rip-off* (negative noun) appear together. Without any rule on the order of appearance of the terms, and so forth, with 50% probability any classifier will produce a misclassification into *positive*. On the other hand, a human coder will classify this text 100% of the times as *negative*.

Despite the rules, which can be highly sophisticated, the problem with ontological dictionaries is that they depend on the language itself, the topic of discussion, the media and so forth. When these contraints are well defined and there is not much noise, for example we are willing to analyze only parliamentary speeches or poems of authors in a given period, then these methods are sufficiently accurate in performing the task. But when the target of the analysis consists of the natural language expressed on social media or the Internet in general where noise is predominant compared to signal, it is quite hard to believe that automatic coding works (even though most computer scientists think it!). Clearly, when automatic tagging is feasible, it is also very convenient as it is reasonably fast and scalable to large amount of data and, most important for science, replicable.

On the other extreme, there is human coding, which is slow but completely data driven rather than assumption driven. As long as the human is able to understand the language and the cultural reference on the language, this coding step is error free to a large extent: irony, metaphoric sentences, jokes and so forth are not problems for humans. Furthermore, algorithms based on supervised human coding are, in fact, language independent as they are completely symbolic in their core. Of course, in some cases bias could be introduced by the coder itself, so more than one coder for each analysis should be used and inter-coder reliability should always be considered carefully.

Human coding also has the benefit of allowing for the classification of complex or long texts where more than one concept is expressed by the writer, as long as the subsequent algorithm is able to accoplish multi-topic classification.

Classification methods

Apart from the dichotomous *supervised/unsupervised* approach, compared to the *scoring* methods, classification algorihtms also split into *individual* versus *aggregated* classification. We now revise some classes of classification methods.

Clustering techniques

Among the *unsupervised* classification techniques we can mention *text mining* methods (Feldman and James 2007; Witten 2004) such as the *cluster analysis*. By *data mining* we mean the set of tools aimed at discovering regularities in the data; by *text mining* we mean the set of techniques able to detect patterns in the texts. Technically speaking, they are the same techniques, just applied to different data, thus they are methodologically indistinguishable. There are, of course, some technical issues like the data preparation process, the notion of distance and so forth which make text mining a relevant field of study *per se*. On final remark is that in *data mining* the information is hidden in the dimensionality of the data, whereas in *text mining* the information is contained in the texts and is visible and transparent though difficult to extract (Hotho et al. 2005). To simplify the exposition, we now focus on the general aspects of *cluster analysis*. This technique is based on the possibility of rearranging observations into homogenous subgroups according to some notion of distance among the data. More precisely, only the notion of *dissimilarity* is sufficient to perform *cluster analysis*. A dissimilarity measure d among two objects A and B, that is $d(A, B)$, is a function which returns zero when it is calculated on the same object, that is $d(A, A) = 0$ for each A (but this does not exclude the possiblity of being zero when evaluated for A and B); it is always non-negative, that is $d(A, B) \geq 0$, and it is symmetric, that is $d(A, B) = d(B, A)$ for all A and B. If in addition, the triangular inequality holds true we call it a *distance* (Gordon 1999). A dissimilarity satisfies the triangular inequality if for any A, B and C, we have $d(A, C) \leq d(A, B) + d(B, C)$. Given a dissimilarity measure d, *clustering* algorithms proceed by grouping (*agglomerative* methods) or splitting (*dissociative* methods) subsequently the whole set of data. If this procedure is sequential, the method is called *hierarchical*. For example an agglomerative hierarchical method is as follows: a first group is formed by taking the closest units in the data. Then each new aggregation occurs, either forming a new group of two units or aggregating a unit to the closest group (according to d) already formed or aggregating two distinct groups. While the distance between two singletons is well defined in general, the distance of a point from a group can be defined in many ways. For example one can take the average distance of the unit to all the units in a group, or the distance of unit from the frontier of the group, or the minimal distance between this singleton and all the units in the group, and so forth. Similar problems exist when the task is to define a distance between two groups. A multitude of rules exist in the literature, and it is clear that each choice produces a different clustering result. To make the analysis less sensible to the initial choice, and hence more robust in the output, usually researchers perform a meta-analysis, which means running many clustering methods and assessing the stability of the solutions. This approach is called *cluster ensembles* (Strehl and Ghosh 2002). Another subtle problem is that the number of clusters is not, in general, the outcome of the cluster analysis but rather a choice of the analyst.

Whichever method of clustering is used, in the end one problem remains: one has to look into the clusters to get some clue of what these clusters mean in term of semantic sense. Supervised clustering methods do exist, and one can always force or help the classification by inserting tagged text into the corpus of data and check *a posteriori* into which cluster these pre-classified data belong. A primer of clustering in textual analysis with dedicated software is Grimmer and King (2011). Keim (2002) also offered a visual approach to *text mining* based on related techniques called *Self-Organizing Maps* (SOM) which helps a lot in summarizing similarites and differences among objects lying in different clusters (Kohonen 2001).

Topic models

Another approach to classification or grouping which is not strictly cluster analysis is called Topic models (Blei 2012; Blei et al. 2003). Topic models apply the LDA (*latent Dirichlet allocation*) method and, as the name suggests, is a way to discover the "topic" of a document. LDA assumes that each text is a mixture of topics. These topics are distributed in the corpus according to some probabilistic distribution (the Dirichlet law). Further, each topic is associated with a sequence of words/stems. The LDA model is, in fact, a Bayesian model with a two-stage data generation process: first a topic is chosen according to a Dirichlet distribution from the set of possible topics, then a set of words/stems is sampled according to multinomial distribution conditionally on the given topic and the result of this data generating process is an observed text. Statistically speaking, the estimation works in reverse: on the basis of the words in a document, the model attaches a probability of belonging to some topic.

Looking at Figure 2.2 we can see the general data generating process. We have K possible topics and each document may contain $\beta_1, \dots, \beta_K$ number of topics, where each β_j is distributed according to a Dirichlet(η) distribution over the words.

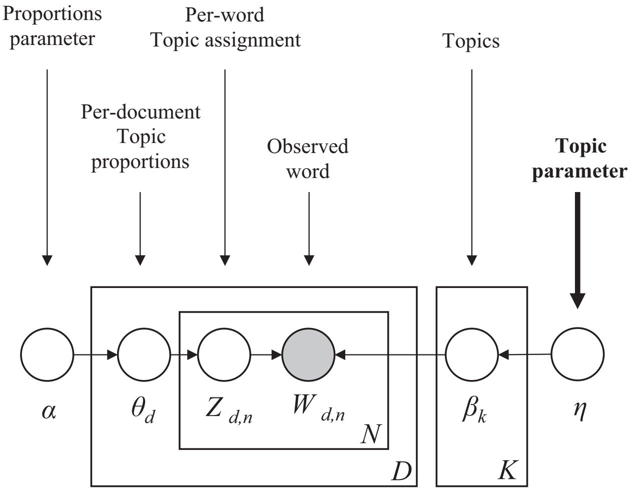

Figure 2.2 The data generating process in topic models

Each *n* word $W_{d,n}$ in a document *d* may belong to more than one topic. The probability of belonging to some topic is given by $Z_{d,n}$ which is the determination of a multinomial distribution. Several hyperparameters exist: θ_d is the proportion of topics per document *d*. This θ_d is also random and distributed according to another Dirichlet distribution or parameter α; and η which is the hyperparameter for the distribution of the β_j. As the only observed data are the words, the goal is to estimate the joint distribution: Pr(topics, proportions, assignments | documents) = $p(\beta, \theta, Z \mid d)$ and then from this assign a topic to each document *d*. This task is far from being easy, and hence several computational steps are needed such as the collapsed Gibbs sampling, the mean field variational methods, expectation propagation and so forth. For full details please refer to (Blei 2012; Blei et al. 2003). If we add a dynamic component, these models take the name of *dynamic multi-topic models* (Quinn 2010). Topic models allow for more general tasks like authorship identification and so forth and, in general, they belong to the supervised techniques as the initial set of topics should be explicated in the process.

Machine learning approach

By machine learning we mean all those techniques that involve *individual* classification of the data in the *test set* given a precoded *training set*. In order to explain why these methods are generally inappropriate in social media analysis, we need more formalism. Assume we have a corpus of *N* distinct texts. Let us denote by $D = \{D_0, D_2, \ldots, D_M\}$ the set of $M + 1$ possible *categories* (i.e. sentiments or opinions) expressed in the texts. Let us denote by D_0 the most relevant category in the data in terms of the probability mass of $\{P(D), D \in D\}$: the distribution of opinions in the corpus. Remark that $P(D)$ is the primary target of estimation in the content of social sciences. We reserve the symbol D_0 to the texts corresponding to Off-Topic or texts which express opinions not relevant with respect to the analysis, that is the *noise* in this framework.

The *noise* commonly present in any corpus of texts crawled from the social network and the Internet in general and should be taken into account when we evaluate a classification method. Suppose that after the stemming step we are left with L stems. The document-term matrix has then N rows and L columns. Let S_i, $i = 1, \ldots, K$, be a unique vector of zeroes and ones representing the presence/absence of the L possible stems. Of course, more than one text in the corpus will be represented by the same vector S_i. Each S_i belongs to the space S of 0/1-vectors of length L, where each element of the vector is 1 if that stem is contained in a text, or 0 otherwise. We denote by s_j, $j = 1, \ldots, N$, the vector of stems associated with the individual text j in the corpus, that is s_j is one of the possible S_i. The set S is theoretically an incredibly large one but, in fact, in each application the cardinality of this set is much lower because the number of actual words and their combinations used in discussion of any given topic are not as large as one might think. So we also assume that K is actual cardinality of S. To summarize, the different dimensions involved satisfy $M \ll L \ll K < N$, where "<<" means "much smaller". Usually, M is around 10, L is at most a few hundred, K millions and N even larger. The corpus of text is ideally divided into two subsets. One of length n, which is called the *training set*, and the remaining one of size N-n, the *test set*.

The texts in the training set are assumed to be already classified without error whilst the data in the training set are not. We denote by (d_j, s_j) the couple that contains the coded value d_j for text s_j. Clearly, d_j is a value in D for the texts in the training set and "NA" for the uncoded texts in the test set. Any machine learning algorithm will try to predict the value of d_j through a model based on the observed vectors s_j. We denote this model by $P(D \mid S)$. Thus, for a text such that $(d_i = NA, s_i)$, the model will estimate the value of d_i according to $P(D = D_m \mid S = s_i)$, $m = 0, 1, \ldots, M$, for example estimating $d_i = D_m$ for the index m such that $P(D = D_m \mid S = s_i)$ reaches the maximal value. Among these methods, we can mention: support vector machines (Cristianini and Shawe-Taylor 2000), random forests (Breiman 2001) and neural networks (Bishop 1995) but also less fancy and much older models like multinomial regression (Engel 1988) or generalized linear models (McCullagh and Nelder 1989) which fall in the same class.

In matrix form, a machine learning algorithm can be written $P(D) = P(D \mid S)P(S)$, where $P(D)$ is a $(M + 1) \times 1$ vector, $P(D \mid S)$ is a $(M + 1) \times K$ matrix of conditional probabilities and $P(S)$ is a $K \times 1$ vector that represents the distribution of S_i over the corpus of texts. The model $P(D \mid S)$ is estimated as $\hat{P}(D \mid S)$ only for the subset of observations in the training set and then the value of $P(D)$ for all the data in the test set is obtained by replacing S with the actual text s_j in $\hat{P}(D \mid S = s_j)$ for s_j in the test set and assigning $d_j = \arg \max D \in D\ \hat{P}(D \mid S = s_j)$. In the naïve nonparametric model, the elements of the matrix $P(D \mid S)$, for example $P(D = D_i \mid S = S_k)$ are estimated by taking the proportion of all texts in the training set that are hand coded as $D = D_i$, which have $s_j = S_k$ as stem vector. Any other model (SVM, etc.) will do essentially the same thing in more sophisticated or more effective ways but the discussion here does not change.

Let us take $i \neq 0$, then for most S_j, these conditional probabilities are zero as the opinion D_i is expressed only for a small subset $S_i \in S$. On the other hand, if we assume that D_0 is the category under which we confine the "*Off-Topic*" texts, then $P(D = D_i|S = S_j) > 0$ for almost all S_i. This means that if n is relatively small, the estimation of $P(D|S)$ will be very poor and, most of the time, the predicted category $D = d$ for a text in the test set will imputed as $d = D_0$, as D_0 is the most frequently observed case in the corpus of texts. As a result, $D = D_0$ will be overestimated and when the aggregation occurs, this strong bias persists so that $P(D)$ will be strongly biased as well.

Another way to understand the previous issues is the following: by D_0 we denote the *noise* or *Off-Topic* category, but the texts belonging to D_0 are, in fact, coming from a different population(s) of texts, that is $D_{noise} = \{D_0\}$, compared to the remaining $D_1, \ldots, D_M$ which are assumed to define the reference population of interest, say $D_{ref} = \{D_1, \ldots, D_M\}$. In this sense, D is the union of two populations $D = D_{noise} \cup D_{ref}$ and this is, in fact, an incorrectly specified reference population or a noise-labeled classification problem (see e.g. Lu et al. 2015). Notice further that D_{noise} itself actually consists of multiple heterogeneous subpopulations as D_{noise} is just the complementary set to D_{ref}. This is why, in essence, any algorithm, including those of machine learning, based on $P(D|S)$ will perform inadequately.[2] In particular, the individual classification error through this approach will remain high and will not vanish due to aggregation because of the large variance in estimates; in contrast, it can easily propagate up to the extent that, in many applications with thousands or millions of texts, one could see the error increasing to 15%–20%. This is clearly quite problematic if one is mainly interested in estimating some type of aggregate measure through the analysis of social media, as occurs precisely with all the studies that want to map (written) opinions to votes. Recently new approaches have been developed (see e.g. Zhang et al. 2015), but we do not consider them here.

Luckily, Hopkins and King (2010), had the idea to change point of view and focus on what can be accurately estimated. Their solution to the problem is as follows:

$$P(D) = [P(S \mid D)^T P(S \mid D)]^{-1} P(S \mid D)^T P(S)$$

under the assumption that the inverse matrix $[P(S \mid D)^T P(S \mid D)]^{-1}$ exists. The previous expression is obtained by the formal inversion of the formula $P(S) = P(S \mid D)P(D)$.

The good news is that in principle, by this approach, $P(S|D)$ can be estimated better than $P(D|S)$ if we have enough coded text for a given category D. Figure 2.3 shows why this is true intuitively. This plot represents the space of Opinion-Stems ($D \times S$). For simplicity, the vectors of stem profiles have been reordered so that the first U supports all the categories from D_0 to D_M, while from $U + 1$ to K supports only the noise category D_0. For example while S_2 supports D_2, D_1 and D_0, any other S_i, for $I > U$, supports D_0 only. Thus, for example $P(S = S_2|D = D_1) > 0$ and

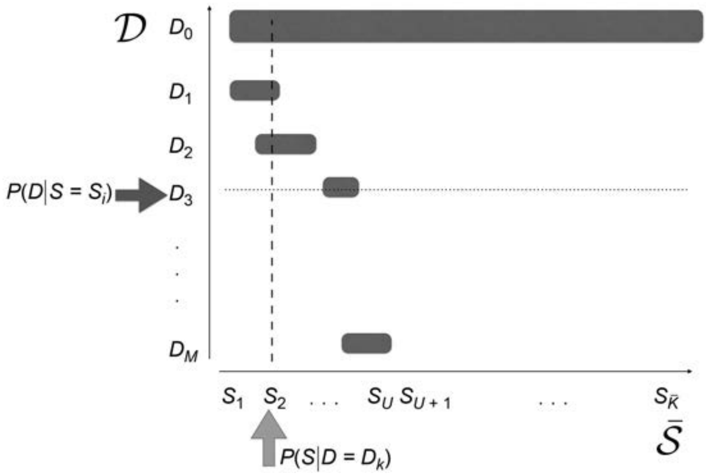

Figure 2.3 Intuitive reason why, when the noise D_0 category is dominant in the data, the estimation of $P(S \mid D)$ is reasonably more accurate than the estimation of counterpart $P(D \mid S)$

$P(S = S_3 \mid D = D_1) = 0$ but $P(D = D_1 \mid S = S_2) = P(D = D_2 \mid S_2)$ is almost zero $P(D = D_0 \mid S = S_2) >> 0$. Roughly speaking, the latter means: most of the time the sequence of stems S_2 is associated with D_0 in the whole document-stem space, while within the category D_2, S_2 has higher probability than for the other categories.

Nevertheless, there is one little issue about this estimation step, as this task involves the estimation over the large number L of stems. This fact makes a direct solution for the problem very difficult. In their seminal paper, Hopkins and King (2010) proposed a workaround to solve this problem. They introduced the algorithm called *ReadMe* based on the repeated random sampling of the number of stems. In ReadMe, the estimate of $P(S \mid D)$ is performed only a random subset of stems over repeated sampling. For each simulation, an estimate of $P(D)$ is obtained, and the results are averaged over many simulations, as is done in a statistical *bagging* approach. One limitation of ReadMe is that the estimates have usually large variance due to the bagging approach and if the number of hand-coded texts per category D is still limited, this may prevent ReadMe to solve the inverse problem in each random sampling step due to bagging. These limitations can be partly attenuated increasing considerably the size of the training set or the size of the subset of stems to use in each bagging replication.

iSA: a fast, scalable and efficient sentiment analysis algorithm

Built on top of the Hopkins and King (2010) inversion formula, iSA (Ceron et al. 2016) is a fast and more accurate implementation of this inverse solution which does not require resampling method and uses the complete length of stems by a simple trick of dimensionality reduction using an idea inherited from Coarsened Exact Matching algorithm (Iacus et al. 2011). The iSA algorithm is as follows:

Step 1 (collapse to one-dimensional vector)

Each vector of stems, for example $s_j = (0, 1, 1, 0, \ldots, 0, 1)$ is transformed into a string-sequence that we denote by C_j = “0110 · · · 01”; this is the first level of dimensionality reduction of the problem: from a $N \times K$ matrix to a one-dimensional vector $N \times 1$.

Step 2 (memory shrinking)

This sequence of zeroes and ones is further translated into hexadecimal notation such that the sequence “11110010” is recorded as λ = “F2” or “11100101101” as λ = “F2D” and so forth. So each text is represented by a label λ of shorter length. When Step 3 is recommended, the string format should be kept.

Step 2b (augmentation, optional)

In the case of nonrandom or sequential tagging of the training set, it is recommended to split the long sequence and artificially augment the size of the corpus as follows. The sequence λ of hexadecimal codes is split into subsequences of length 5, which corresponds to 30 stems in the original 0/1 representation. For example suppose we have the sequence λ_j = “F2A10DEFF1AB4521A2” of 18 hexadecimal symbols and the tagged category d_j = D3. The sequence λ_j is split into $4 = \lfloor 18/5 \rfloor$ chunks of length 5 or less: λ^1 = “aFEA10”, λ^2 = “bDEFF1”, λ^3 = “cAB452” and λ^4 = “d1A2”. At the same time, the d_j is replicated (in this example) four times. The same applies to all sequences of the corpus. This method produces a new dataset which length is four times the original length of the dataset, that is 4N. When Step 2b is used, we denote iSA as iSAX (where “X” stands for sample size augmentation). The estimation of $P(\lambda \mid \mathrm{D})$ is as easy as an instant tabulation over the n labels λs (or 4n after data augmentation) of the training set which has complexity of order n (or 4n) and N (or 4N) for $P(\lambda)$.

Step 3 (quadratic programming)

Whether or not Step 2b has been applied, this step solves the inverse problem into a single (QP) run. The whole set of equations becomes $P(D) = [P(\lambda \mid D)^T P(\lambda \mid D)]^{-1} P(\lambda|D)^T P(\lambda)$. The assumption is that the inverse matrix $[P(\lambda \mid D)^T P(\lambda |D)]^{-1}$ exists but now $P(\lambda|D)$ is more dense than the original counterpart $P(S \mid D)$ as there is higher probability to find similar shorter subsequences λ_j than in the full sequence S_j in the data.

Step 4 (bootstrapping)

This step allows for the estimation of the standard errors of the estimates of $P(D)$. Conversely, for example ReadMe, iSA is not a *bagging* method, that is it uses all the available stems in a single QP step, eventually after the optional data augmentation procedure. This allows for the application of a standard bootstrap algorithm (see, e.g. Efron and Tibshirani 1993) that, in this case, provides accurate and unbiased estimates of the mean and variance of $P(D)$. Standard bootstrap is performed by resampling the rows of the document-term matrix after dimensionality reduction, providing an extremely fast solution to the bootstrapping problem. In the following examples, we show how to obtain bootstrap standard errors with R.

The properties of iSA alone and in comparison with other machine learning tools as well as the ReadMe approach have been shown in full detail in Ceron et al. (2016) to which readers are redirected. Here we present an application of iSA through the R package name iSAX.

The iSAX package for the R statistical environment

The iSAX package is available for academic use only through the GitHub system at http://github.com/blogsvoices/iSAX. In order to use the package, the iSAX package has to be installed into R. We assume that readers are familiar with this statistical software (R Core Team 2016). Once in the R Console, if not present, the user has to install the devtools package as follows:[3]

```
R> install.packages("devtools")
```

When the installation is completed, the iSAX package can be installed from within R typing the following commands:

```
R> library(devtools)
R> install_github("blogsvoices/iSAX")
```

If nothing goes wrong, the iSAX package is ready to be used typing the following command:

```
R> library(iSAX)
```

There is no need to reinstall the package; from the next R session, iSAX will be ready to run.

We now show an easy application of the iSA workflow through the iSAX package. As a working example, we use the benchmark database of the Internet Movie Database (IMDb) which is a corpus of 50,000 hand-coded movie reviews prepared by Mass et al. (2011). Each film has been classified according to scale from 1 to 10. Levels 1 to 4 are considered *negative* reviews; Levels from 7 to 10 correspond to *positive* reviews. Neutral reviews are not present in this example data. In our analysis, this corpus is further divided into a training set and a test set, each of 25,000 cases. To test the validity of iSA, we disregard the hand-coded information in the test set and try to predict the aggregated distribution. The data are loaded into R as follows:

```
R> data(mov.train)
R> data(mov.test)
```

We can check the length of each corpus:

```
R> length(mov.train)
[1] 25000
R> length(mov.test)
[1] 25000
```

We can inspect the content as follows:

```
R> str(mov.train[[1]])
List of 2
$ content: chr “Bromwell High is a cartoon comedy.
It ran at the same time as some other programs
about school life, such as \”Teachers\”. My 3”|
__truncated__
$ meta :List of 10
..$ author : chr(0)
..$ datetimestamp : POSIXlt[1:1], format: “2015-05-09
10:51:59”
..$ description : chr(0)
..$ heading : chr(0)
..$ id : chr “0_9.txt”
..$ language : chr “en”
..$ origin : chr(0)
..$ sentiment : chr “positive”
..$ set : chr “train”
..$ stars : chr “9”
..- attr(*, “class”)= chr “TextDocumentMeta”
- attr(*, “class”)= chr [1:2] “PlainTextDocument”
“TextDocument”
```

The information about the review is contained in the metadata in the field stars. So we can extract it as follows and check the distribution of the reviews in the training set and in the test set:

```
R> unlist(lapply(mov.train, function(x)
meta(x)$stars)) -> D.train
prop.table(table(D.train))
```

```
D.train
1 10 2 3 4 7 8 9
0.20400 0.18928 0.09136 0.09680 0.10784 0.09984
0.12036 0.09052
R> unlist(lapply(mov.test, function(x) meta(x)$stars))
-> D.test
prop.table(table(D.test))
D.test
1 10 2 3 4 7 8 9
0.20088 0.19996 0.09208 0.10164 0.10540 0.09228
0.11400 0.09376
```

as well as the distribution of the whole corpus, which is our target:

```
> prop.table(table(c(D.train,D.test)))
1 10 2 3 4 7 8 9
0.20244 0.19462 0.09172 0.09922 0.10662 0.09606
0.11718 0.09214
```

We now prepare the corpus for the analysis. The iSAX package has a function called prep.dat which also performs the stemming. For full details, we refer to the iSAXpackage documentation.

```
R> corpus <- c(mov.train, mov.test)
R> ocome <- prep.data(corpus, th=0.95,verbose=TRUE)
Phase1: Cleaning up . . .
stripping white spaces . . .
Phase2: stemming . . .
Phase3: bin2hexing . . .
```

The data have been prepared for iSA. The binhexed feature vector is contained in the slot S and the document-term matrix in the slot dtm. We can inspect both

```
S <- ocome$S
R> head(S,2)
1
“000000080020220000000080201000000004000200082800800008
000400000000020910000000008840000020008”
2
“0100000000820201080000000000000054004000001001000800 00
400080000000000100008020029000000000000”
R> dtm <- ocome$dtm
R> dim(dtm)
[1] 50000 364
```

The stemming has left a total of 364 stems. The first entries of the document-term matrix look as follows:

```
> dtm[1:5,1:8] # first five rows, first 8 columns
Terms
Docs absolut act action actor actual almost along also
1 0 0 0 0 0 0 0 0
2 0 0 0 0 0 0 0 1
3 0 1 0 0 0 0 0 0
4 0 0 0 0 0 0 0 0
5 0 0 1 0 0 0 0 1
```

We now construct the vector of tags by using the real tags for the training set and NA for the test set:

```
R> tr <- c(D.train, rep(NA, length(D.test)))
R> table(tr)
tr
1 10 2 3 4 7 8 9
5100 4732 2284 2420 2696 2496 3009 2263
```

We can verify that the data contain 25,000 missing data:

```
R> summary(factor(tr))
1 10 2 3 4 7 8 9 NA's
5100 4732 2284 2420 2696 2496 3009 2263 25000
```

We can now proceed with iSA as follows:

```
R> idx <- which(is.na(tr)) # index of the NA data
R> Stest <- S[idx] # selection of the test set from
the corpus
R> Strain <- S[-idx] # selection of the training set
R> D <- tr[-idx] # the tag from the training set only
R> set.seed(123) # for replication purposes
R> iSAOut <- iSA(Strain, Stest, D, nboot=100) # the
iSA algorithm
iSAX . . .
bootstrapping . . . please wait
Elapsed time: 6.52 seconds
```

We can now inspect the outcome which is contained in the slot btab of the outcome value of iSA:

```
R> iSAOut$btab
Estimate Std. Error z value Pr(>|z|)
```

```
1 0.20023858 0.007816733 25.61666 9.950740e-145
10 0.21798596 0.005606032 38.88418 0.000000e+00
2 0.09354359 0.005970674 15.66717 2.536094e-55
3 0.10114215 0.007645071 13.22972 5.910648e-40
4 0.10447341 0.009252117 11.29184 1.439930e-29
7 0.08640438 0.005590567 15.45539 6.939333e-54
8 0.11077500 0.006794878 16.30272 9.439321e-60
9 0.08543693 0.005223087 16.35756 3.842435e-60
```

In the previous outcome, the Estimate and Std.Error columns are, respectively, the bootstrap mean and standard errors of the *P(D)* estimates. *P* values are evaluated under the Gaussianity assumption which holds true asymptotically.

Finally, we can compare the iSAX estimate with the true target distribution:

```
R> mat <- cbind(iSAOut$btab[,”Estimate”], prop.
table(table(c(D.train,D.test))))
R> colnames(mat) <- c(“iSA”,”True”)
R> round(mat,3)
iSA True
1 0.200 0.202
10 0.218 0.195
2 0.094 0.092
3 0.101 0.099
4 0.104 0.107
7 0.086 0.096
8 0.111 0.117
9 0.085 0.092
We can calculate the Mean Absolute Error as follows
R> mae <- mean(abs(mat[,1]-mat[,2]))
R> mae
[1] 0.006777923
```

which is pretty low in this example. The same example run for ReadMe, random forests and support vector machines produced the following results taken from Table 2.3 in Ceron et al. (2016):

The full R scripts to replicate the results are available as additional material to the published paper.

Table 2.3 Comparison of individual and aggregated classifiers

Algorithm	*RF*	*SVM*	*ReadMe*	*iSA*	*iSAX*
MAE	0.060	0.002	0.041	0.002	0.007
Execution Time	953.1s	5289.8	95.2s	2.1s	8.0s

A working example from raw data to sentiment and opinions: the Donald Trump data

These data have been collected for an undergraduate political science class at the University of Milan. It is a “toy example” just to show the students and readers of this book the necessary steps to run an analysis with iSA. We warn readers (and the students!) that no scientific conclusion can be drawn from this tutorial analysis. The dataset is a sample of about 2% of all the tweets mentioning the official account of Donald Trump “@realDonaldTrump”, on dates 7–13 June 2016, written in English and coming from the US. Data have been collected through Twitter API also specifying language and origin of tweets.

The data have been later hand coded by 10 undergraduate students and made available on the GitHub repository for the R package iSAX at http://github.com/blogsvoices/iSAX as a CSV (Comma Separated Value) file just for the aforementioned purpose. Data consists of 28,741 tweets, with about 400 hand-coded tweets along three dimensions: sentiment (positive/neutral/negative) and two sets of opinions: “Why negative?” and “Why positive?”. In a case where Trump is mentioned incidentally or to mean things different from the electoral competition, the corresponding tweet is marked as OffTopic. As mentioned earlier, it is essential to always mark properly this category.

As in the previous part, the symbol “R>” is the prompt on the R Console and should not be typed. We first download the data as follows:

```
R> library(iSAX)
R> x <- read.csv(“https://raw.githubusercontent.
com/blogsvoices/iSAX/master/Trump.csv”,
stringsAsFactors=FALSE)
R> str(x)
‘data.frame’: 28741 obs. of 4 variables:
$ text : chr “This Video Will Get Donald Trump
Elected - if it Goes Viral fb.me/3jHfL0E7g” “RT @
MOVEFORWARDHUGE: DIARY: well, this’s awkward! I know,
I’ll blame Trump 4 terrorist attacks and bluster! TONS
of bluster AND”| __truncated__ “RT @PuyaFiel: Hispanic
Trump supporter tells Protester \”Go back to Mexico!\”
But Media doesn’t want to show this pic.twitter.c”|
__truncated__ “La Raza Council Cancels Pro-Muslim,
Anti-Trump Event After Massive Terrorist Attack ln.is/
dailycaller.co. . . via @dailycaller” . . .
$ Sentiment : chr “neutral” “positive” “positive”
“neutral” . . .
$ WhyPos : chr NA “general endorsement” “foreign
policy” NA . . .
$ WhyNeg : chr NA NA NA NA . . .
```

We now organize the tweets in a corpus using the tm package which is automatically loaded by iSAX

```
R> corpus <- VCorpus(VectorSource(x$text))
```

and we prepare the data for iSA algorithm. The function prep.data performs stemming and other preprocessing steps and returns a vector of features S and the corresponding document-term matrix dtm.

```
R> ocome <- prep.data(corpus,verbose=TRUE, th=0.995)
Phase1: Cleaning up . . .
stripping white spaces . . .
Phase2: stemming . . .
Phase3: bin2hexing . . .
```

and we inspect the results:

```
R> str(ocome)
List of 4
$ S : Named chr [1:28741] “000000000000000000410000008
00000000000000000000000000000000000000000000000000800
c001000” “00020002000000000000000000000000020000100000
00000000000000000000000000000001008000004000” “000100
00000000000080000000000000100000000000008400000000000
0200000000200084008000100000” “000a000000000000000000
00000000000000000000000000000000000000000000000000001
000010000000” . . .
..- attr(*, “names”)= chr [1:28741] “1” “2” “3” “4”
. . .
$ dtm : num [1:28741, 1:345] 0 0 0 0 0 0 0 0 0 0 . . .
..- attr(*, “dimnames”)=List of 2
.. ..$ Docs : chr [1:28741] “1” “2” “3” “4” . . .
.. ..$ Terms : chr [1:345] “account” “actual” “agre”
“alreadi” . . .
$ train : NULL
$ th : num 0.995
```

We can notice that the stemming step has left 345 stems in the document-term matrix and the vectors *S* contain the rows of the document-term matrix in hexadecimal notation as explained in the iSA algorithm.

We now investigate the Sentiment of the whole data and later the motivations of positive/negative sentiment, that is the opinions. First we need to extract the coded values and separate the set in training and test set. The training set corresponds to the values of the variable *Sentiment* which are not missing.

```
R> train <- !is.na(x$Sentiment)
negative neutral offTopic positive
191 178 10 103
R> table(x$Sentiment)
R> table(train)
train
FALSE TRUE
28259 482
```

So we have 482 coded tweets for the Sentiment dimension. We now select the features *S* for the training and test set and synch the coded values with the training set for iSA,

```
R> Strain <- ocome$S[which(train)]
R> Stest <- ocome$S[-which(train)]
R> length(Strain)
[1] 482
R> length(Stest)
[1] 28259
D <- x$Sentiment[train] # the D vector for iSA
prop.table(table(D))
D
negative neutral offTopic positive
0.39626556 0.36929461 0.02074689 0.21369295
```

and the latter is the empirical distribution of *D* within the training set. Finally, we run iSA. For replication purposes we set the seed of the random number generator which is used in the bootstrapping step:

```
> set.seed(123)
> outSent <- iSA(Strain,Stest, D)
iSAX . . .
bootstrapping . . . please wait
Elapsed time: 0.58 seconds
> round(outSent$btab, 5)
Estimate Std. Error z value Pr(>|z|)
negative 0.40094 0.03803 10.54356 0.00000
neutral 0.41706 0.03299 12.64023 0.00000
offTopic 0.01912 0.01030 1.85623 0.06342
positive 0.16288 0.02856 5.70296 0.00000
```

From the previous table we can see the estimated values for *P*(*D*) and the corresponding standard errors and *p* values of the asymptotic *t* test. It looks like the Off-Topic category is not predominant (indeed, the *p* value is larger than 5%) in

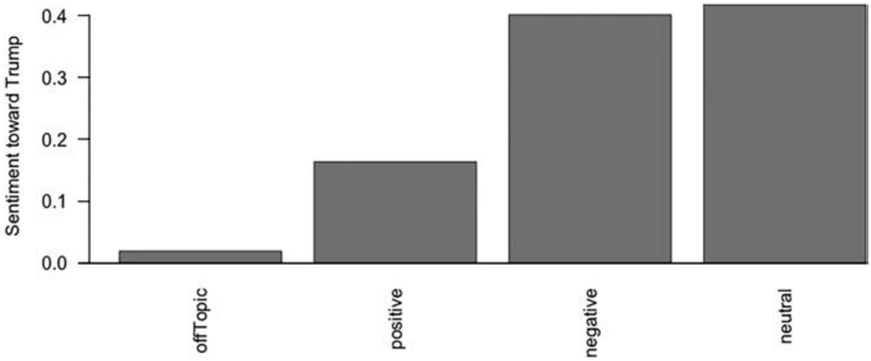

Figure 2.4 The estimated sentiment toward Donald Trump

this set and there is a general negative or neutral sentiment toward the candidate whilst only 16% is positive.

We can plot the data as follows:

```
R> tabSent <- as.table(sort(outSent$btab[,"Estim
ate"]))
R> par(mar=c(8,5,1,0))
R> barplot(tabSent,col="red", ylab="Sentiment toward
Trump",las=2)
```

The results are in Figure 2.4.

We now proceed through the same steps and look at the reasons behind a positive sentiment.

```
R> train <- !is.na(x$WhyPos)
R> table(x$WhyPos)
foreign policy general endorsement honesty
9 50 6
leadership offTopic
2 10
R> table(train)
train
FALSE TRUE
28664 77
```

In this case, 77 coded values as available, some with an extremely low number of valid tagging. We should expect some instability with general classifier, but iSA can still provide an answer.

```
R> Strain <- ocome$S[which(train)]
R> Stest <- ocome$S[-which(train)]
```

```
R> length(Strain)
[1] 77
R> length(Stest)
[1] 28664
R> D <- x$WhyPos[train]
R> prop.table(table(D))
D
foreign policy general endorsement honesty
0.11688312 0.64935065 0.07792208
leadership offTopic
0.02597403 0.12987013
R> set.seed(123)
R> outWhyPos <- iSA(Strain,Stest, D)
iSAX . . .
bootstrapping . . . please wait
Elapsed time: 0.66 seconds
> round(outWhyPos$btab, 5)
Estimate Std. Error z value Pr(>|z|)
foreign policy 0.09913 0.04124 2.40361 0.01623
general endorsement 0.52081 0.06499 8.01314 0.00000
honesty 0.08824 0.03230 2.73158 0.00630
leadership 0.08277 0.03676 2.25181 0.02433
offTopic 0.20905 0.04017 5.20367 0.00000
```

At 5% level, all items of the vector *P*(*D*) are statistically significant, but clearly not at 1%. We can now plot the results as before. They are visible in Figure 2.5.

```
R> tabWhyPos <- as.table(sort(outWhyPos$btab[,”Estim
ate”]))
R> par(mar=c(11,5,1,0))
R> barplot(tabWhyPos,col=“green”, ylab=“Why
Positive?”,las=2)
```

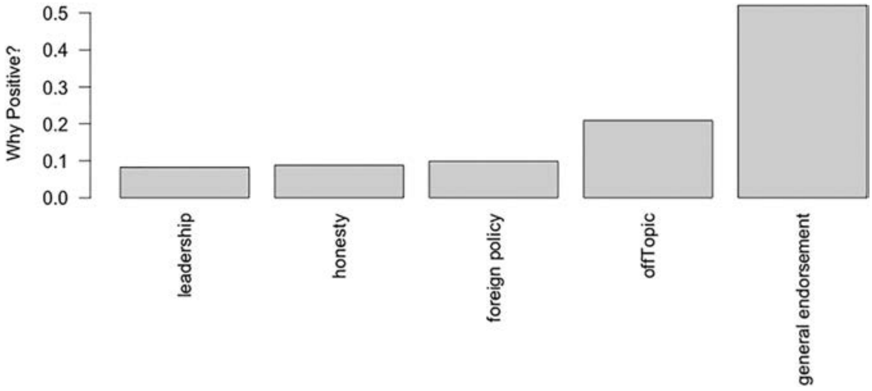

Figure 2.5 Reasons for positive sentiment toward Donald Trump

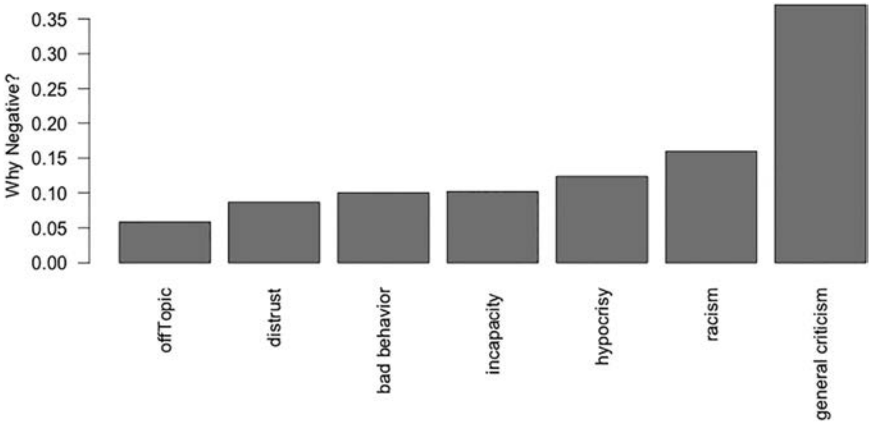

Figure 2.6 Reasons for negative sentiment toward Donald Trump

We now end the analysis with the reasons behind the negative sentiment. The code is run without comments and the results are plot in Figure 2.6.

```
R> train <- !is.na(x$WhyNeg)
R> table(x$WhyNeg)
bad behavior distrust general criticism
6 7 53
hypocrisy incapacity offTopic
8 13 10
racism
27
R> table(train)
train
FALSE TRUE
28617 124
R> Strain <- ocome$S[which(train)]
R> Stest <- ocome$S[-which(train)]
R> length(Strain)
[1] 124
R> length(Stest)
[1] 28617
R> D <- x$WhyNeg[train]
R> prop.table(table(D))
D
bad behavior distrust general criticism
0.04838710 0.05645161 0.42741935
hypocrisy incapacity offTopic
```

```
0.06451613 0.10483871 0.08064516
racism
0.21774194
R> set.seed(123)
R> outWhyNeg <- iSA(Strain,Stest, D)
iSAX . . .
bootstrapping . . . please wait
Elapsed time: 0.80 seconds
R> round(outWhyNeg$btab, 5)
Estimate Std. Error z value Pr(>|z|)
bad behavior 0.09988 0.01919 5.20511 0.00000
distrust 0.08624 0.02787 3.09444 0.00197
general criticism 0.37054 0.05294 6.99959 0.00000
hypocrisy 0.12364 0.02764 4.47407 0.00001
incapacity 0.10147 0.02789 3.63837 0.00027
offTopic 0.05854 0.02176 2.69105 0.00712
racism 0.15968 0.03082 5.18098 0.00000
R> tabWhyNeg <- as.table(sort(outWhyNeg$btab[,”Estim
ate”]))
R> par(mar=c(8,5,1,0))
R> barplot(tabWhyNeg,col=“red”, ylab=“Why
Negative?”,las=2)
```

In this case, all the entries of vector *P*(*D*) are statistically significant.

Notes

1 By “hot” we mean a word or a hashtag that is frequently used in discussions about a given topic.

2 In the literature, when applying automated data classification a precision of 80% in classifying an individual post in a given “true” category is considered satisfactory (Mostafa 2013; Oliveira et al. 2015; Yoon et al. 2013).

3 The “R>” should not be typed, as it represents the R Console prompt.

References

Benoit, K., and Laver, M. (2003) ‘Estimating Irish party positions using computer wordscoring: The 2002 elections’, *Irish Political Studies*, 18(1): 97–107.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press.

Blei, D. (2012) ‘Probabilistic topic models’, *Communications of the ACM*, 55(4): 77–84.

Blei, D., Ng, A., and Jordan, M. (2003) ‘Latent Dirichlet allocation’, *Journal of Machine Learning and Research*, 3: 993–1022.

Breiman, L. (2001) ‘Random forests’, *Machine Learning*, 45(1): 5–32.

Ceron, A. (2015) ‘The politics of fission: An analysis of faction breakaways among Italian parties (1946–2011)’, *British Journal of Political Science*, 45(1): 121–139.

Ceron, A., Curini, L., and Iacus, S.M. (2016) ‘iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content’, *Information Sciences*, 367–368: 105–124.

Cristianini, N., and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press.

Curini, L., Hino, A., and Osaki, A. (2013) ‘Measuring party competition from legislative speeches: Analysis of Japanese parliamentary debates, 1953–2011’, Presented at ECPR General Conference, Bordeaux, 4–7 September.

Dave, K., Lawrence, S., and Pennock, D.M. (2003) ‘Mining the peanut gallery: Opinion extraction and semantic classification of product reviews’, *Proceedings of WWW 2003*, Budapest, Hungary, 20–24 May 2003.

de Boeck, P., and Wilson, M. (2004) *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer.

Efron, B., and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC, CRC Press.

Engel, J. (1988) ‘Polytomous logistic regression’, *Statistica Neerlandica*, 42(4): 233–252.

Feldman, R., and James, S. (2007) *The Text Mining Handbook*. Cambridge, UK: Cambridge University Press.

Gordon, A.D. (1999) *Classification* (2nd Ed.). Boca Raton, FL: Chapman & Hall/CRC.

Grimmer, J., and King, G. (2011) ‘General purpose computer-assisted clustering and conceptualization’, *Proceedings of the National Academy of Sciences*, 108(7): 2643–2650.

Grimmer, J., and Stewart, B.M. (2013) ‘Text as data: The promise and pitfalls of automatic content analysis methods for political texts’, *Political Analysis*, 21(3): 267–297.

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

Hotho, A., Nurnberger, A., and Paaß, G. (2005) ‘A Brief Survey of Text Mining’. *LDV Forum*, 20(1): 19–62.

Iacus, S.M., King, G., and Porro, G. (2011) ‘Multivariate matching methods that are monotonic imbalance bounding’, *Journal of the American Statistical Association*, 106: 345–361.

Jurafsky, D., and Martin, J.H.. (2009) *Speech and Natural Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Keim, D.A. (2002) ‘Information visualization and visual data mining’, *IEEE Transactions on Visualization and Computer Graphics*, 7(2): 100–107.

Kohonen, T. (2001) ‘Self-Organizing Maps’, Springer Series in Information Sciences, Vol. 30. Berlin, Springer.

Laver, M., Benoit, K., and Garry, J. (2003) ‘Extracting policy positions from political texts using words as data’, *American Political Science Review*, 97(2): 311–331.

Liu, B. (2006) *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, NY (1997 Mitchell, TM Machine Learning) New York: WCB/McGraw-Hill.

Lu, Z., Wu, X., and Bongard, J.C. (2015) ‘Active learning through adaptive heterogeneous ensembling’, *IEEE Transactions on Knowledge and Data Engineering*, 27(2): 368–381.

Maas, D., Pham, H., and Ng, P. (2011) ‘Learning word vectors for sentiment analysis’, *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, Portland, OR, 19–24 June 2011.

McCullagh, P., and Nelder, J. (1989) *Generalized Linear Models* (2nd Ed.). Boca Raton: Chapman and Hall/CRC.

Mostafa, M.M. (2013) ‘More than words: Social networks’ text mining for consumer brand sentiments’, *Expert Systems With Applications*, 40(10): 4241–4251.

Oliveira, D.J.S., de Souza Bermejo, P.E., and dos Santos, P.A. (2015) ‘Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls’, *Journal of Information Technology & Politics*. doi: 10.1080/19331681.2016.1214094.

Pang, B., and Lee, L. (2004) ‘A sentimental education: Sentiment analysis using subjectivity’, *Proceedings of the ACL-04 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 25–26 July.

Pang, B., Lee, L., and Vaithyanathan, S. (2002) ‘Thumbs up?: Sentiment classification using machine learning techniques’, *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing,* Philadelphia, PA, 6–7 July 2002.

Pennebaker, J.W., Boyd, R.L., Jordan, K., and Blackburn, K. (2015) *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin. doi: 10.15781/T29G6Z

Quinn, K. (2010) ‘How to analyze political attention with minimal assumptions and costs’, *American Journal of Political Science*, 54(1): 209–228.

R Core Team (2016) R: ‘A language and environment for statistical computing’, R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/

Slapin, J., and Proksch, S.-O. (2008) ‘A scaling model for estimating time-series party positions from texts’, *American Journal of Political Science*, 52(3): 705–722.

Strehl, A., and Ghosh, J. (2002) ‘Cluster ensembles–A knowledge reuse framework for combining multiple partitions’, *Journal of Machine Learning Research*, 3: 583–617.

Witten, I.H. (2004) *Text Mining, Practical Handbook of Internet Computing*. Boca Raton, FL: CRC Press.

Yoon, S., Elhadad, N., and Bakken, S. (2013) ‘A practical approach for content mining of tweets’, *American Journal of Preventive Medicine*, 45(1): 122–129.

Zhang, J., Wu, X., and Sheng, V.S. (2015) ‘Active learning with imbalanced multiple noisy labeling, *IEEE T*’, *Cybernetics*, 45(5): 1081–1093.

3 Nowcasting and forecasting electoral campaigns

Evidence from France, the United States and Italy

Summary of electoral forecasts made through SASA

In the current and in the next chapter we show how the information available on social media can be successfully used to *nowcast* and *forecast* electoral campaigns. We focus in particular on five different examples. In Chapter 5, these forecasts, along with many other predictions produced by different scholars (and through different methods), are then used to assess the accuracy and the reliability of social media-based electoral forecasts.

The list of studies presented and analyzed in the present and in the following chapter includes the main elections that took place in France, the United States and Italy between April 2012 and December 2013. In detail, we have considered the first and the second round of the French presidential elections (April and May 2012), the presidential election in the United States (November 2012), the first and the second round of the primary election of the Italian center-left coalition (November and December 2012), the Italian general election (February 2013) and the selection of the leader of the main Italian party, that is the Democratic Party (December 2013). As a result, the cases differ both in terms of the type of election considered (a national election devoted to select the head of state versus an election aimed to select the leader of a political coalition running in the next national election or the secretary of a party versus a multiparty election aimed to select a government) and in the type of competition involved (a "single-issue" election in which the preference eventually expressed by Internet users involved mainly a choice among two options versus an election in which Internet users could choose to express a preference among a larger number of potential viable targets). Moreover, in the United States, the rate of penetration of Twitter,[1] that is the social network on which we focus here, is considerably larger than in Italy or France,[2] implying an Internet community not necessarily identical in the two cases. Such dissimilarities are clearly important for exploring the robustness of our results and for controlling the potentiality of a method that, for a number of different reasons explained subsequently, seems to be an advance compared to other available methods.

Before turning to these examples, however, some first words are in order. First, all the analyses that we discuss were obtained by employing the SASA approach discussed in the previous chapter (either the algorithm ReadMe – in our first attempts of electoral forecasts – or the iSA one – in the later studies).[3]

Second, all the predictions were carried out using, as already noted, Twitter data. Twitter is a microblogging social media platform on which users post messages in real time; these are called "tweets" and have a maximum length of 140 characters (Silva et al. 2014). Broadly speaking, there are several social media sites that can be analyzed. The advantage of Twitter, which makes it so popular in the literature on social media analysis, is that most posts by users ("tweets" in Twitter jargon) are freely accessible if collected in real-time, contrary to other mostly closed social networks such as Facebook, Instagram, and so forth (Vargo et al., 2014). Another important feature of Twitter as a data source is that the 140-character limitation on tweets forces users to be more concise and consequently to express their views in a more obvious way, which makes the processing of the data rather effective (Kontopoulos et al. 2013; Oliveira et al. 2015). Moreover, through analyzing Twitter is also possible to geo-localize the origin of the tweet, therefore permitting a more fine-grained analysis (as we will see subsequently).[4] Thus, Twitter can be considered a rich and varied source of data for the application of sentiment analysis and opinion mining (Khan et al., 2014).

Third, all the results were made publicly available *before* the elections. In this sense, they can be strictly considered as *real* forecasts. Notice that the population of tweets that we collected consists of all the texts posted during the pertinent time period that include in their content at least one of a set of keywords (the name of the political leaders/parties covered by each of our analysis as well as the most popular hashtags characterizing each candidate/party's electoral campaign). With respect to predictions based on Twitter data, our first analyses were made downloading data through the Twitter application programming interface (API).[5] It is well known that the limitation imposed to the Twitter API makes it hard to gather the whole population of data from this source (Morstatter et al. 2013), hence we relied on some particular techniques in order to partially overcome these limits (following Sampson et al. 2015) and to maximize the amount of tweets downloaded. However, in recent years, Twitter has increased the limitations imposed on the API. As a consequence, to produce the prediction based on a large enough number of comments (closest as possible to the whole population of tweets), we decided to download tweets by way of buying data from an authorized Twitter data provider.

Fourth, we always employed the same rules to map tweets into votes. More in detail, we deemed that a post expresses a real intention to cast a vote in favor of a candidate/party only if at least one of the following three conditions is satisfied:

1 The post includes an explicit statement related to the willingness to vote for a candidate/party,
2 The post includes a statement in favor of a candidate/party together with a message or a hashtag connected to the electoral campaign of that candidate/party,
3 The post includes a negative statement opposing a candidate/party together with a message or a hashtag connected to the electoral campaign of a rival candidate/party.

Considering a positive statement plus a campaign message or a hashtag, and not simply a generic positive statement, permits one to focus on signals that are more "costly" in terms of self-exposition and that are, therefore, more credible (for this point, see the sizeable literature on signaling games: Banks 1991). In contrast, the third condition allows one to reduce the arbitrariness in the "supervised" stage of the analysis. This also applies to a (largely) two-candidate case, such as the US presidential race. For example if a tweet says "Do not vote for Romney", this does not necessarily imply that the person who wrote that post will then vote for Obama. He or she could decide to vote for a third candidate or to abstain. In a multiparty race, of course, this problem is even more significant. Returning to the previous example, a hypothetical tweet such as "Do not vote for Romney. #fourmoreyears" would be counted, according to our classification, as a vote in favor of Obama, given that *#fourmoreyears* has been one of most largely used hashtags supporting Obama's electoral campaign.[6]

Similarly, we counted all the retweets (i.e. the diffusion by a Twitter user of a message posted by another user) that satisfy the previous conditions as a ''vote'' for a candidate/party. Although retweeting, strictly speaking, does not imply the production of new information, it implies that someone else thought a communication was valuable for herself (Jensen and Anstead 2013). On the other hand, if it is true that the act of retweet does not necessarily imply an ''endorsement'' by the user who retweets, it is also true that when the retweet includes a text in which an intention to vote a given candidate/party is clearly expressed or where an identifiable hashtag connected to a candidate/ party is presented, it becomes a costly act, exactly for the same reasons already noted previously. As a consequence, it should happen only when a Twitter user shares to a large extent the content of the tweet and the underlying connected vote.

Fifth, per each analysis that is discussed in this chapter (as well as in the next chapters) we provide a separate box in which we highlight a set of interesting information related to data collection, data processing and data release. In detail, we report: *Period* (the range of dates used in the analysis), *Data Source* (the list of social media considered), *Data Gathering* (explaining whether data collection was performed through API or buying data from a data provider), *Keywords* (the list of words used in the search query to perform data collection), *No. comments* (the number of texts downloaded and analyzed), *SASA Method* (to discriminate between ReadMe and iSA), *Weights* (to explain if any form of weighting has been applied), *Election day* (the day when the election took place), *Forecast release* (to detect which prediction was provided daily and to report the date of the last forecast) and *Reference* (a list of active websites reporting the ex ante prediction or any other source that can witness that the prediction was made ex ante).

Finally, the distinction between nowcasting and forecasting (two words we have used up to now without any further comments) deserves here a brief discussion. The term *nowcasting* is, in fact, different from the idea of *forecasting* (i.e. predicting the future) although the two concepts are not completely unrelated (see Lewis-Beck et al. 2011 for the first use of the term *nowcasting* in the election forecasting literature). Forecasting implies anticipating something that has yet to come; nowcasting implies capturing something that is happening in this exact moment, that is it implies moving from predicting the future, to predicting the present. This apparent paradox (how can you use the word *predict* for something that has already happened?) is easily explainable.

Nowcasting is the ability of releasing immediately the information that can be extracted from the data, that is the exercise of reading, through the lenses of a model, the flow of data released in real time (Banbura et al. 2013). For example one could use social media data to nowcast hidden real-time quantities of interest (Lampos and Cristianini 2012; Lansdall-Welfare et al. 2012) such as the mood of a population or the presence of a flu epidemic (as we discussed in Chapter 1).

Indeed, official statistics or survey polls, with their natural reporting lags, may not always be timely indicators (due to the effort in sampling and data collection). Moreover, survey responses often involve recall of past events and experiences and are elicited at discrete moments chosen by researchers. On the contrary, social media content is posted continuously and often contemporaneously with what is being analyzed. This is especially relevant with respect to the voting intentions of citizens that can suddenly change after an episode related to the electoral campaign. In light of this, the idea of nowcasting the elections implies that we can be in the conditions to effectively monitoring the sign of a trend, or its switch, in a faster way if compared to what can be done through survey polls. Through that, real-time online and social media data allow for continuous analysis as events unfold and temporally-granular post-event analysis critical to isolating the impact of key sub-events (Diaz et al. 2016). This produces better insightful inferences, as we will see, about which episode of the campaign actually affected the degree of support toward a party or a candidate.

Hollande versus Sarkozy: monitoring the 2012 French presidential election

Our first attempt to use social media data in order to nowcast the evolution of the campaign and to forecast the final result was made in April and May 2012 focusing on the first (22 April) and second round (6 May) of the French presidential elections.

2012 French presidential election (1st round)	
Period	15–21 April 2012
Data Source	Twitter
Data Gathering	API
Keywords	hollande, sarkozy, sarkò, bayrou, melenchon, joly, mélenchon, dupont, aignan, election, presidentielle, élection, présidentielle, poutou, arthaud, le pen
No. comments	158,584
SASA Method	ReadMe
Weights	No
Election day	22 April
Forecast release	22 April (morning)
Reference	http://voicesfromtheblogs.com/2012/04/22/presidenziali-francesi-su-twitter/

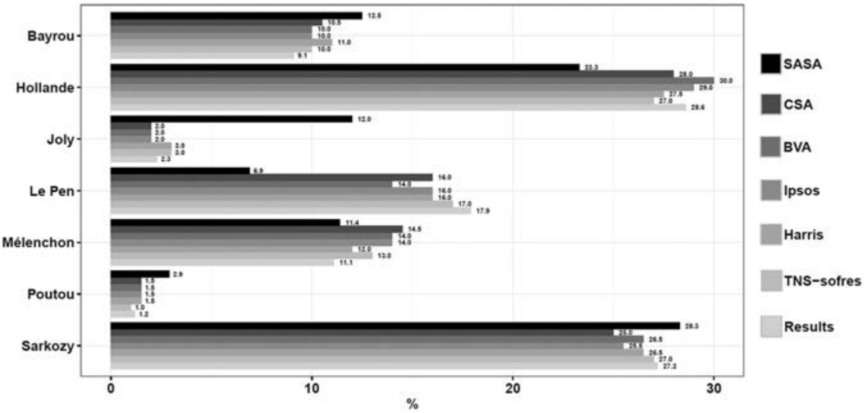

Figure 3.1 First round of the 2012 French presidential election: vote share predictions according to electoral surveys and SASA by presidential candidates (SASA)

To start with, we investigated the opinions related to the first round by downloading and analyzing over 158,000 tweets published in the last 7 days before the election. Seven candidates were taken into consideration: Bayrou, Hollande, Joly, Le Pen, Mélenchon, Poutou, and Sarkozy. Our prediction was published in the morning of Election Day.[7] This very first attempt looked promising, given that the MAE of our prediction, measured by comparing the SASA estimates with the actual results of the polls, for all seven candidates taken into account, was equal to 4.65. Figure 3.1 compares our estimates with those provided by survey polls. Although the MAE of the surveys was lower (always lower than 2), we noticed that for many candidates our prediction was precisely in line with both survey data and the actual results. This is particularly true for Sarkozy, but also for Mélenchon, whose SASA prediction was very precise while the performance of this left-wing candidate was overestimated by any other source of data (including survey data, count of mentions or automated sentiment analysis). What is more, the SASA prediction proved to be much more precise than predictions based on automated techniques of sentiment analysis, whose MAE was higher (6.70).[8]

To delve into the predictive power of SASA, we also decided to investigate the second round of the election. Once again, we decided to focus only on the last days before the election, assuming that – given the importance of the election – a large part of the everyday conversations, both online and offline, would be devoted to discussing the themes of the campaign and the reason for choosing one or the other candidate.

2012 French presidential election (2nd round)	
Period	27 April–5 May 2012
Data Source	Twitter
Data Gathering	API
Keywords	hollande, sarkozy, presidentielle, election, elysee, Objectif2012, RadioLondres, UMP election, AvecSarkozy, VoteHollande, PS election, FH2012, NS2012, AuRevoirNS, jevoteNS, votonsSarko, Limoges, sarko, SarkoCaSuffit, hollandesarkozy, France2012, ledebat9
No. comments	263,948
SASA Method	ReadMe
Weights	No
Election day	6 May
Forecast release	Daily; last release: 5 May (afternoon)
Reference	http://voicesfromtheblogs.com/2012/05/02/presidenziali-francesi-su-twitter-secondo-turno/

Indeed, the election attracted the attention of Twitter users and, in the last 10 days (before Election Day), more than 263,000 tweets were downloaded and analyzed to detect the voting intentions expressed online. By doing that, we were able to monitor, day by day, the evolution of the expressed preferences. Such opinions and their variation over time can be put in relation with the events of the campaign such as debates and rallies, proposals and promises or scandals and other exogenous events.

Between 27 April and 5 May we ran eight different analyses. The results are reported in Figure 3.2. As shown, Hollande was almost always the front-runner (in 7 observations out of 8). His advantage, however, widened or shrank according to the events of the campaign.

On the first day, the two were almost head-to-head and the socialist candidate was only slightly ahead. On 28 April, to the contrary, we observed a peak in the voting intentions in favor of Hollande. The reason was related to two scandals involving the incumbent President Sarkozy that emerged on that day. First, Sarkozy was accused of having obtained illicit funds from Gadhafi's regime in order to finance Sarkozy's campaign in 2007. Second, the media reported news related to an alleged illegitimate act of spying carried out by the intelligence agency against Dominique Strauss-Kahn to weaken this foremost socialist politician and to prevent him from running into the election.

On the next day, however, the Libyan government officially denied the news about illicit funds provided by Gadhafi to Sarkozy. Accordingly, the Sarkozy staff attacked the media, accusing them of having intentionally lied because they support Hollande. On the same day, the entrance of Dominique Strauss-Kahn into the campaign produced a backlash effect for the socialist candidate. As Strauss-Kahn took part to an electoral meeting organized by the Socialist Party, the supporters of

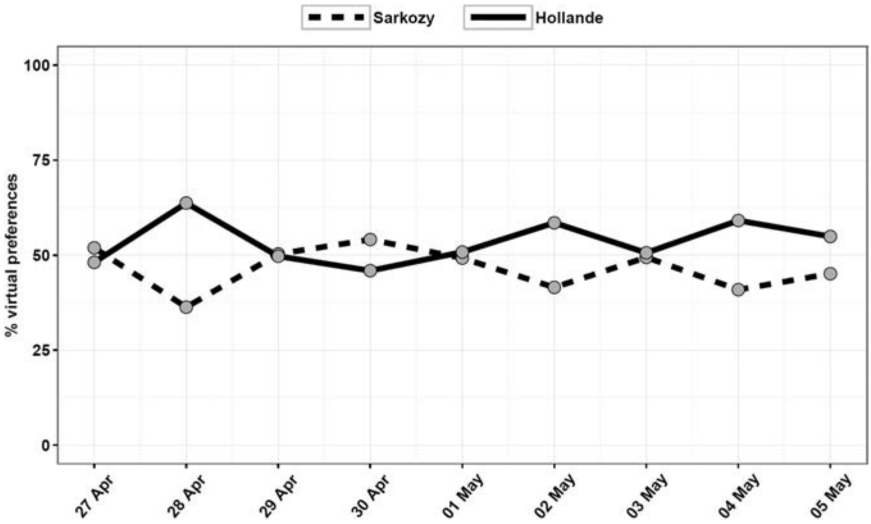

Figure 3.2 Flow of preferences expressed on Twitter during the electoral campaign for the second round of the 2012 French presidential election (SASA)

Sarkozy exploited this episode to attack Hollande by strongly criticizing the bad reputation of Strauss-Kahn on valence issues.

Jointly, these two events allowed Sarkozy to rank first in the voting intentions expressed on 30 April. This advantage, however, vanished on the very next day due to a mistake made by Sarkozy. On 1 May, in fact, he declared that rather than the International Workers' Day we should celebrate only the "real work". Such declaration sounded insulting to hundreds of workers and voters, and its effect was to decrease the share of support toward Sarkozy. An emblematic comment stated: "*on May 1st N.Sarkozy celebrates the 'real work'. On May 6th we will offer him the real holidays.*" The race was very close and for this reason the television debates held on 2 May became potentially decisive. On that day, almost 70,000 online comments were downloaded and analyzed; such rate was approximately three times higher if compared to the previous days.

The debate was a success for Hollande ("*Tonight you understood: Hollande refused to hold 3 debates to respect the dignity of Sarkozy*"), while the performance of Sarkozy was poor as noted by several Twitter users ("*The enemy of François Hollande? Is Nicolas Sarkozy. And that of Nicolas Sarkozy? Is Nicolas Sarkozy again.*"). The advantage gained by Hollande thanks to the debate was pretty huge, and it was confirmed also on the following days. As such, it could have really made the difference in winning or losing the election. Conversely, the endorsement in favor of Hollande made by the centrist candidate Bayrou, leader of the MoDem (Democratic Movement), did not seem to produce strong effects. We noticed this when looking at the reasons provided by users that expressed their intentions to vote for Hollande or Sarkozy. On the one hand, those talking about the debate were overall more enthusiastically supportive of Hollande. On the other, voters talking about the endorsement made by Bayrou were equally split between those who thought that such endorsement would have positive effects on Hollande's vote share and those arguing that it would not be useful for him. For instance some comments appreciated the endorsement ("*Congratulation to Bayrou for his brave choice to vote Hollande. Given the right-wing drift of Sarkozy this is a good sense choice*") and wished to thank Bayrou ("*Thanks Bayrou, for the first time in history, the centre is leaning toward the left*"). Conversely, others highlighted his incongruity ("*Bayrou criticizes the platform of Hollande but he is voting for him, in fact, he doesn't hold any certainty*") and thought ("*the choice of Bayrou will help Sarkozy*").

Our final prediction made through Twitter predicted the victory of Hollande, assigning to him 54.9% of voting intentions, which is a value close to that of survey polls. In the last days before the election, in fact, surveys assigned to Hollande a share of votes between 52% and 53.5%. Hollande,, indeed, won the election with the 51.64% votes.

Based on this, our MAE was equal to 3.26 whereas the MAE of survey polls ranged between 0.4 and 4.4 and, on average, it was equal to 2.09. Therefore, our prediction was comparable to that of traditional surveys. Notice that, averaging out all the predictions made by survey polls from 22 April, the expected vote share of Hollande (53.7%) was identical to the average of the predictions based on Twitter (that are reported in Figure 3.3).

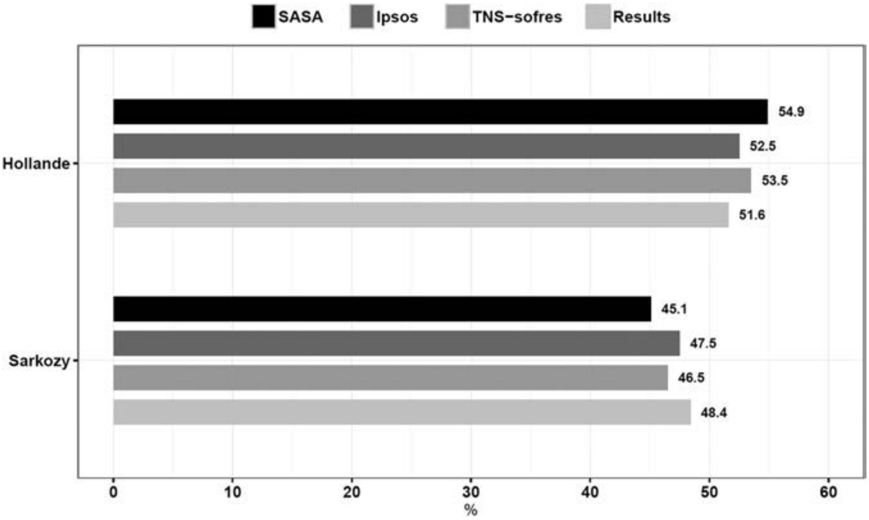

Figure 3.3 Second round of the 2012 French presidential election: vote share predictions according to electoral surveys and SASA by presidential candidates

The 2012 US presidential election: too close to call?

Due to the large diffusion of social media and to the usage of social networking sites as a room for enacting fundraising and e-campaigning (e.g. Cogburn and Espinoza-Vasquez 2011), the United States context has attracted the interest of scholars willing to forecast elections through social media. A number of published or unpublished papers has tried to assess the final outcome of legislative, local, primary or presidential elections (Barclay et al. 2015; Choy et al. 2012; Contractor et al. 2015; DiGrazia et al. 2013; Gayo-Avello 2011; Gordon 2013; Huberty 2015; Jensen and Anstead 2013; MacWilliams 2015; Mejova et al. 2013; Murthy 2015; Nooralahzadeh et al. 2013; Shi et al. 2012; Topsy Lab 2012; Washington et al. 2013).

These attempts date back to the 2008 presidential election (Gayo-Avello 2011), which was the first campaign in the Era of Web 2.0. The attempt to predict the outcome of 2008 US presidential election using Twitter data, however, was unsuccessful. It is worth recalling that Twitter was created only a couple of years before, in 2006. Arguably, the time was not ripe yet to exploit such information (Gordon 2013).

Just to give an idea, that study focused on tweets written between 1 June 2008 and 11 November 2008 that were geo-located in one of the following states: California, Florida, Indiana, Missouri, North Carolina, Ohio, Texas. Montana was planned to be included as well, but it was discarded due to lack of data, as only 817 tweets geo-located there were found.

As such, that analysis was based solely on the tweets published by 20,000 users and, in the last 6 months before the elections, only 250,000 tweets were downloaded and analyzed. Therefore, in each state, the average number of tweets per month varied between 2,000 and 8,000 and only in California we found 17,000 tweets per month.

These figures are very distant from the millions of tweets written four years later, in 2012. In fact, as we will see, in the last weeks before the 2012 presidential election, more than 50 million tweets were collected and analyzed. What is more, in the swing states (discussed later) the average number of geo-located tweets – per day – was around 10,000 units (i.e. roughly 30 times more than in 2008).

In 2012 the Twitter environment was more widespread, becoming a prominent arena for political communication. Accordingly, a number of studies has looked at Twitter data to analyze the race for the White House, starting with the primary elections held by the Republican Party (Jensen and Anstead 2013; Shi et al. 2012).

2012 US presidential election	
Period	28 September–6 November 2012
Data Source	Twitter
Data Gathering	API
Keywords	romney, obama, vote obama, vote romney, obama2012, mitt2012, obamabiden, romneyryan, teaparty, romneyryan2012, tlot, tcot, gop, GOP2012, rnc, Virgil Goode, Jill Stein, Gary Johnson, #ForAll, #Forward, dems, democrats, 4moreyears, p2, topprog, fivemoredays

(*Continued*)

2012 US presidential election	
No. comments	Around 50 million
SASA Method	ReadMe
Weights	No
Election day	6 November
Forecast release	Daily; last release: 6 November (24.00, CET)
Reference	RAI 1, Porta a Porta (talk show broadcast by Italian National Television); http://www.corriere.it/esteri/speciali/2012/elezioni-usa/; http://sentimeter.corriere.it/2012/11/07/elezioni-e-social-network-ancora-una-volta-la-rete-anticipa-il-voto/

In view of that, from 28 September to 6 November we monitored the "*voting intention*" (according to the rules defined earlier) expressed on Twitter toward the two main candidates: Barack Obama (Democratic Party) and Mitt Romney (Republican Party) as well as to voting intentions toward others. In this lapse of time, we estimated the electoral preferences of American voters, from Alabama to Wyoming, on a daily basis by analyzing more than 50 million tweets, a bit more than 1 million tweets per day. Each datum is calculated as a moving average along those 7 days (following what suggested in O'Connor et al. 2010).[10] These estimates were published daily on the online version of the main Italian newspaper, *Corriere della Sera*.[11] The results are summarized in Figure 3.4.

Beside the electoral preferences for the presidential candidates, by means of SASA we also measured the content of tweets to gauge which were the most

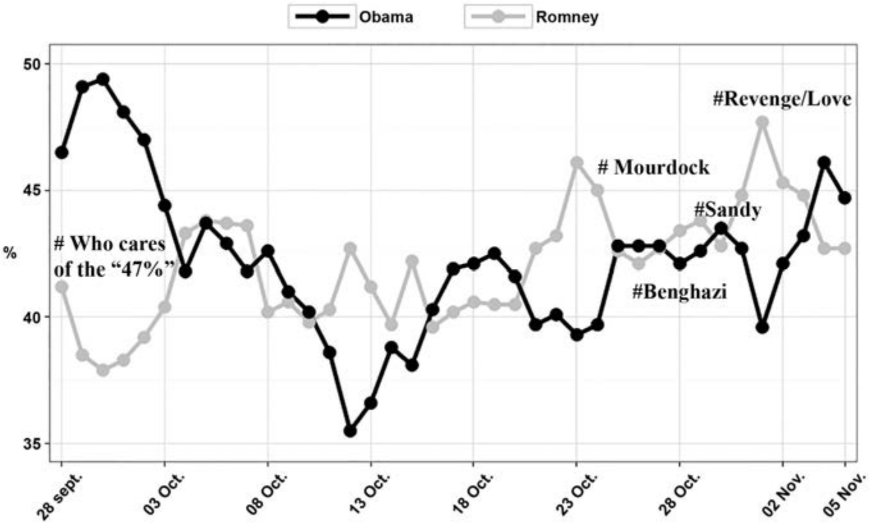

Figure 3.4 2012 US presidential election: daily social voting intentions according to SASA

Table 3.1 Temporal evolution of the main topics of the US presidential campaign according to social media conversations

Period	*Main topics*
28 Sept.–7 Oct.	Economy (unemployment and taxation); Welfare (including Obamacare)
8 Oct.–13 Oct.	Economy (unemployment and taxation); Civil rights (gender, immigration, abortion)
14 Oct.–28 Oct.	Economy (unemployment and taxation); Civil rights (gender, immigration, abortion); Foreign policy (Benghazi)
29 Oct.–2 Nov.	Civil rights (gender, immigration, abortion); Environment and energy
3 Nov.–6 Nov.	Generic appeals to voters

discussed topics that drove Twitter conservations (an exercise of what we have called in Chapter 2 as "opinion analysis"). The state of the economy was the most relevant theme for most of the campaign, though online conversations also focused on civil rights, health policies (Affordable Care Act, also known as Obamacare) and immigration. In this regard, Table 3.1 describes the evolution of conversations over time, summarizing – per each time period – the most discussed campaign topics. We will return later on this table.

On average, we can note that the fluctuation of the preferences expressed online closely followed the main events that happened during the electoral campaign. From Hurricane Sandy to "Benghazi-gate", the whole campaign received a strong echo on social media, particularly during the three television debates that involved the two main candidates, Obama and Romney.

When looking at Figure 3.4, we notice that Obama and Romney were leading in the online preferences for almost half of the time each. In the 40 days of the analysis Obama was leading 21 times while Romney was on top on the remaining 19 days. This is rather similar to what was highlighted by survey polls: both sources of data on public opinion revealed how the race was close and how the two main competitors were trying to surpass each other during the campaign. In fact, the (rather large) victory of Obama was the consequence of a final rush, even though the incumbent president was clearly leading over Romney for the whole month of September.

In September, after the Democratic National Convention and thanks to the several gaffes made by Romney, the support for Obama was huge and he received close to 50% of voting intentions. In particular, the video of a private party meeting that went viral (also on Twitter) dramatically damaged the Republican candidate. In that video footage, Romney attacked the "47%" of Americans, blaming them for being parasitic citizens who do not pay income tax and only rely on the subsidies of the government to survive. Indeed, after this gaffe few Twitter users felt comfortable in publicly expressing support for him.

Due to these events, the surveys predicted a boring campaign with a clear winner, but the analysis of social media told a slightly different story. In this regard, the wind started to change (on Twitter) in October, when Romney was able to reduce the distance from the Democratic candidate. Immediately since the first days of October, the trend of online conversations appeared different. In those days, the most discussed topics (see Table 3.1) were related to the economy, including taxes, deficit and the costs of Obamacare. Due to the non-brilliant state of the economy, the incumbent president was punished in terms of approval as these themes entered into the agenda of the campaign. Conversely, the supporters of Romney managed to have their voice heard in online conversations and criticized Obama for the poor results in terms of unemployment.

On Twitter, the gap already narrowed, and in this regard the first television debate made the difference, as Romney overcame the incumbent president in the voting intention of Twitter users afterward. The first debate was held on 3 October in Denver, Colorado. Once again, the state of the economy was the main topic addressed in the debate and, on this issue, Obama was put in the corner. The performance of Obama was poor, particularly if compared to that of Romney, who was active and aggressive: "*there must be consideration of invoking the Mercy Rule somewhere. Mitt Romney just smoking him,*" a tweet said. This episode recalled a moment during the Republican National Convention in which the actor Clint Eastwood debated on stage with an empty chair that represented a silent Obama.

In the debate, Romney addressed the concerns of the moderate, independent and undecided voters and became perceived as a credible leader. In evaluating the debate, Twitter users openly judged Romney as the clear winner. Just to quote a couple of comments, users said "*Romney is owning this debate. He is living in that area right of center*" or analogously "*Wow. Not a Mitt fan, but I'm finding a LOT I agree with in his debate*" and finally, some comments also used ironic statements to express opinions: "*If Romney keeps this up, Obama is gonna vote for him.*"[12] On that evening, in fact, 70.5% of tweets claimed that Romney had won while only 29.5% argued that Obama was the winner. Remarkably, this evaluation was in line with the opinions recorded by survey polls (according to which the ratio between Romney and Obama, in the debate, was 73–27. See Table 3.2).[13]

Table 3.2 Reaction to the presidential debates during the US electoral campaign according to survey polls and Twitter (SASA)

	Twitter sentiment	*Survey polls*	*MAE(average: 3.0)*
First Debate (3 October 2012)	Obama: 29.5 Romney: 70.5	Obama: 27.2 Romney: 72.8	2.3
Second Debate (16 October 2012)	Obama: 53.1 Romney: 46.9	Obama: 54.1 Romney: 45.9	1.0
Third Debate (22 October 2012)	Obama: 60.3 Romney: 39.7	Obama: 54.5 Romney: 45.5	5.8

It is also interesting to notice that, due the relatively slow pace of polling, traditional survey polls revealed a trend similar to that of social media opinions only in the days following the first television debate while the conversations of online public opinion seemed to anticipate such trend. This highlights what we previously noted about the ability of social media analysis to nowcast any "momentum" during an electoral campaigning (see next section on this point).

Thanks to the boost provided by the debate, the share of voting intentions for Romney oscillated between 40% and 45%. Those in favor of Obama declined over time, even when the policy agenda focused on post-materialist issues like gender, immigration and civil rights that could have helped him to mobilize the consent of minorities and increase his rate.

The second debate, held at the Hofstra University of New York, represents another turning point that arrested the loss of support for Obama. The incumbent president radically changed his strategy and, rather than presenting himself as a moderate and nonpartisan figure who aimed to unify the country, Obama started to fight. Once again, the debate involved economic topics. *Deficit* and *middle class* were among the most cited words online as well as the retrospective judgment on the "record" obtained by the economic policy of Obama. He also attacked Romney, talking about social policy and, in fact, the words *women*, *black* and *immigration* were mentioned in a wide number of tweets. Indeed, Obama was the winner of this second debate according to the 53.1% of tweets while only 46.9% of comments were favorable to Romney (a survey related to the debate produced almost exactly the same results).[14]

Thanks to that, Obama surpassed Romney also in the share of voting intentions. This lead, however, did not last despite the fact that Obama also dominated the third debate, held on 22 October. This third debate focused on foreign policy; *troops* and *Muslim* were among the words most used. Obama took advantage of his role as Commander in Chief and claimed credit for the *killing* of Osama Bin Laden (another widely discussed topic), which made the United States a safer place. Indeed, the hashtag *#strongerwithobama* was widely used during the debate. The experience of Obama in foreign policy was quite huge and Romney could not attack the incumbent president on this. Accordingly, he adopted a defensive strategy, agreeing with Obama on many topics. In fact, the only attack made by Romney against Obama during this debate proved to be unsuccessful and generated a powerful backlash effect. Romney criticized Obama for having cut military expense for the Navy, but the Obama answered with a joke that became one of the symbols of the campaign: Obama replied to Romney, arguing that the military expenses were cut indeed, but he cut expenses related to "horses and bayonets" as nowadays there are many other ways to secure the country. The hashtag *#bayonetsandhorses* went viral and was repeatedly used by citizens and activists until the end of the campaign. It goes without saying that Obama won this last debate according to both Twitter (60–40) and to survey polls (55–45).[15]

The similarities between the online sentiment and the output of offline surveys in the evaluation of the performance of the two candidates during the three debates are quite impressive (see Table 3.2). This can be a signal that, notwithstanding the results of other studies (Mitchell and Hitlin 2013), the reaction of online public opinion to political events presents many analogies with that of the overall public opinion and goes in the same direction. Table 3.2 reports the comparison between the sentiment of live tweeting comments. It was measured through SASA and published online immediately after the debate.[16] The sentiment was measured to assess, according to the tweets written during the debate, which of the two candidates was to be considered the winner. These results were compared with those of three CNN/ORC International polls of people who watched the debate, asking respondents who was the winner of the quarrel (data reported net of answers like "unsure" and "don't know").[17] As we can see, the two sources produced very similar results. In the last column, we report the mean absolute error of the estimates, which in this case expresses the distance between sentiment and survey. This distance was on average equal to 3 and was remarkably lower in the first and second debates.

Sadly for Obama, a political scandal emerged in those days, halting the increase in his voting intentions. In fact, a few days after the second debate, some generals of the US Army blamed Obama for being responsible for the so-called Benghazi-gate. According to them, Obama did not exploit all the information made available to him and did not do his best to prevent the killing of the US ambassador to Libya. Furthermore, Obama was accused of not telling the truth when answering to questions about Benghazi-gate. To lie can be very a damaging choice (Callander and Wilkie 2007; Davis and Ferrantino 1996), and this was even more true for Obama, as only a few days before these accusations he launched a new slogan, arguing that "*This election is about trust.*" The involvement in Benghazi-gate seemed to suggest to voters that they could not really *trust* Obama, and this effect was quite long lasting, as it grabbed the attention of media and voters for several days: between the 23rd and the 28th of October, the hashtag *#benghazi* remained one of the most popular on Twitter, and this effect produced a drop in the Obama's voting intentions.

This affair could have been problematic for Obama's chances to be reelected, as Romney kept leading by 5 points (45–40). However, another event suddenly erupted, counterbalancing the *#benghazi* effect and damaging the Republican candidate. This was related to a gaffe made by Richard Mourdock, a Tea Party member running for the Senate in Indiana for the Republican party. In a public statement, Mourdock said that he opposed abortion (for religious concerns) in case of rape, because even such a birth comes as God's will. Exactly like Benghazi-gate, such statement had a huge echo on Twitter, and Romney's party started to be increasingly considered as too close to extremist Christian groups. Although Romney was not the author of such gaffe, his share of voting intentions decreased, and the two candidates were head-to-head again.

A few days later, another "exogenous" event jumped into the campaign: Hurricane Sandy. Obama managed to take advantage of it. He canceled his campaign events, wore a military jacket and visited New York and New Jersey, two states hit by the hurricane. Neither was a swing state, and both were considered as safely leaning toward the Democratic Party. Therefore, the visit made by Obama was not perceived as an electoral strategy. Nevertheless, such behavior strengthened the perception of Obama as a strong and inclusive leader, up to the point that two foremost Republican politicians, Governor Chris Christie of New Jersey and Mayor Michael Bloomberg of New York, thanked Obama for his effort. His ready reaction to the hurricane was put in contrast with some old declarations delivered by Romney, who suggested cutting public expenditure for the civil protection, and this point boosted the online support for Obama.

In that last week, policy issue and exogenous events disappeared from the agenda of the campaign. Both offline political rallies and online conversations focused just on the final slogans and appeals to the voters. On the one hand, Obama demanded *#fourmoreyears*. On the other, Romney answered by launching a sort of countdown and arguing that only *#fivemoredays* were left for the Obama presidency. The last attempts to mobilize voters were made by both candidates: Obama called to vote for *#revenge* while Romney's appeal was about voting for *#love* of the country. Finally, on Election Day, the hashtag *#stayinline* was used by Democratic activists to explain to Democratic voters that those who were already in line were entitled to vote, even beyond the closing times of polling stations. These messages, which were widely circulated in the key swing states such as Virginia, Florida and Ohio could, indeed, have contributed to the victory of Obama.

Nevertheless, the share of online voting intentions highlighted a positive trend toward Obama in the very last days of campaign. Accordingly, our final prediction made on 6 November forecasted a victory for Obama in the popular votes with a clear and safe margin of 3.5%. As the real gap in the share of votes was 3.9%, our forecast proved (surprisingly) to be more accurate than those made by traditional survey polls that on average assigned only a narrow margin in favor of Obama (+0.7) claiming that the race was "*too close to call*".[18]

Beside the popular vote, we also tried to predict the results in the swing states, that is those where the race is usually very close and a few thousands of votes can alter the balance between the candidates and the outcome of the whole presidential election.[19]

In 2012, the surveys focused on 11 swing states, and among these they considered Florida, Ohio and Virginia as the main battlegrounds. We, therefore, paid attention to those races. In Table 3.3, we report the gap between Obama and Romney in each state, according to three different measures. The first one (labeled SASA) consists in the method of sentiment analysis discussed so far. For the swing states estimates, we replicated our method, considering the pools of tweets geo-tagged in each state only. The second (R) is the gap displayed on 6 November on Realclearpolitics.com, which recorded the average of the survey polls issued on the last week before the election. The third one (V) represents the actual gap between the two candidates after votes have been counted. Then we display the difference between the forecasts (made either through sentiment analysis or surveys) and the actual votes. Finally, we highlight what prediction has been the best one according to the ability to correctly predict the winner. When both sentiment analysis and survey polls predicted the same winner, we discriminated by measuring the difference between the expected and the actual gap.

Overall, analyzing social media correctly predicted the winner in 9 out of 11 swing states, Colorado and Pennsylvania being the only two exceptions. Furthermore, in a plurality of states (7 against 2), our data (HK) proved to be more accurate than the average of polls (these states are Florida, Iowa, Virginia, Nevada, New Hampshire, Michigan, Wisconsin), while in the remaining two swing states (Ohio and North Carolina) the different forecasts (social media versus surveys) performed in a similar manner.

Table 3.3 2012 US presidential election: accuracy of the predictions

State	*Gap (SASA)*	*Gap (R)*	*Gap (V)*	*|SASA-V|*	*|R-V|*	*Best prediction*
Popular Vote	Obama +3.5	Obama +0.7	Obama +3.9	0.4	3.2	*SASA*
Florida	Obama +6.1	Romney+1.5	Obama +0.9	5.2	2.4	*SASA*
Ohio	Obama +2.9	Obama +2.9	Obama +3.0	0.1	0.1	=
Virginia	Obama +3.5	Obama +0.3	Obama +3.9	0.4	3.7	*SASA*
Colorado	Romney +1.3	Obama +1.5	Obama +5.4	4.1	3.0	R
Iowa	Obama +4.8	Obama +2.4	Obama +5.8	1.0	3.4	*SASA*
Nevada	Obama +3.3	Obama +2.8	Obama +6.7	3.4	3.9	*SASA*
New Hampshire	Obama +3.8	Obama +2.0	Obama +5.6	1.8	3.6	*SASA*
North Carolina	Romney +3.0	Romney +3.0	Romney +2.0	1.0	1.0	=
Michigan	Obama +5.5	Obama +4.0	Obama +9.5	4.0	5.5	*SASA*
Pennsylvania	Romney +2.5	Obama +3.8	Obama +5.4	2.9	1.6	R
Wisconsin	Obama +7.4	Obama +4.2	Obama +6.9	0.5	2.7	*SASA*

Note: Comparison between SASA method and survey polls (R) estimates with the actual results (V).

The most interesting results concern the three main battlegrounds: Ohio, Virginia and Florida. In Ohio, both tools predicted similar results. On the contrary, in Virginia our data proved able to catch the pro-Obama trend that emerged in the last days (and in the last hours) when the Democratic staff mobilized the partisan voters (even during the electoral night they pushed voters to stay in line by means of Twitter messages). The same happened in Florida, where our prediction claimed a victory for Obama, with a safe margin. Eventually these results could be explained by our ability to measure the voting intention of the Hispanic voters, who could be less likely to answer to survey polls.

This goes back once again to the specific method we employed to analyze social media. In fact, while the SASA method was remarkably able to catch the voting intention of US citizens, other methods adopted to analyze social media in the same occasion were less successful. Returning to a point already made throughout the book, it is interesting to note that at the beginning of the electoral campaign Obama had almost 16.8 million followers on Twitter, while Romney had not even hit 600,000. The same happens when we look, for instance at the number of Facebook friends (Barclay et al. 2015): on 6 November, Obama had 32 million likes on his official Facebook profile while Romney had only approximately one third that number (12 million).

Despite such (huge) disparity, our results underlined a different story, that was not only remarkably in line with the actual votes, as we have discussed, but that also illustrated a social media support for the two main competitors that was much more volatile compared to what we could have expected by looking at the number of followers only (see Figure 3.1). In this sense, our results confirmed that the number of Facebook friends or Twitter followers on their own is largely misleading as predictors of election outcomes (see Cameron et al. 2013 on this point).

This happens also because Twitter users are often about divided between those who follow leaders they agree with and those who also follow political figures they disagree with (see Parmelee and Bichard 2011).

The same unsatisfactory prediction could have happened if we had counted the number of mentions related to the different candidates running in the 2012 United States presidential election (Nooralahzadeh et al. 2013; Washington et al. 2013). Once again this result is not surprising, given that the sheer number of mentions related to a candidate gives just a measure of the notoriety in the web of such candidate (either for positive *or* negative reasons), without any necessary connection with his or her (expected) voting share. Techniques of automated sentiment analysis were sometimes more useful in the 2012 US elections, though their accuracy remains lower compared to the SASA method. For instance Twindex, the Twitter Political Index developed by Topsy and Twitter itself, estimated the day ahead of the election a wide margin for Obama in terms of positive sentiment compared to Romney (74% against 59%).[20] Washington et al. (2013), on the other hand, found contrasting evidence since they were able to get somewhat accurate estimates only when using one commercial algorithm that provided a social media marketing platform (for details, see Washington et al. 2013), while their results get much worse when applying other dictionary-based techniques.[21] Finally, between September and November, Contractor et al. (2015) mixed 37 million tweets (which is only a sample of the total number of tweets, due the rate limits restrictions in the API) with Gallup surveys, and they predicted a gap of 0.6 points between the two candidates, that is 3.3 points lower than the actual one. Analogously, Choy et al. (2012) showed that automated sentiment analysis can wield good predictions even though this accuracy approaches the one reached by the SASA technique only when the results are weighted by some census information, such as preexisting party affiliations of the American voters.[22]

Table 3.4 provides an evaluation of the different techniques in the 2012 US presidential election case. To allow a comparison between these studies we considered

Table 3.4 Comparison of the accuracy of Twitter forecast made through mentions, automated sentiment analysis and SASA method (2012 US presidential election, popular vote)

Method	*Source*	*MAE*
Mentions	Nooralahzadeh et al. 2013	8.39
Mentions	Washington et al. 2013	17.90
Sentiment Analysis	Topsy Lab 2012	3.63
Sentiment Analysis	Washington et al. 2013	1.80
Sentiment Analysis	Washington et al. 2013	16.00
Sentiment Analysis + Surveys	Contractor et al. 2015	1.65
Sentiment Analysis	Choy et al. 2012	1.29
Sentiment Analysis + Weights	Choy et al. 2012	0.47
SASA		*0.02*

only the two main candidates, Obama and Romney, and normalized the Twitter results and the popular vote data accordingly. We then measured the MAE of each prediction. Table 3.3 simply shows the difference, in terms of MAE, between the SASA and the other automated techniques discussed earlier. The table illustrates that the SASA sentiment analysis clearly performs better than the alternatives, at least in the 2012 US case.

Surveys versus sentiment: who leads in the 2012 US presidential election?

In the previous section, we showed how analyzing data from Twitter can provide, under some given circumstances, not only a precise prediction of the final outcome of an election (a point on which we return in greater detail in Chapter 5), but also an accurate description of the evolution of an electoral campaign. This last point brings us naturally to the possibility of nowcasting. As discussed earlier, nowcasting an electoral campaign means to be in the conditions to monitoring the sign of a trend, or its switch, in a faster way compared to survey polls through the availability of real-time social media. In this respect, the 2012 US presidential race discussed in the previous section is an interesting case study thanks to the public availability through Realclearpolitics.com of a rather long time series of electoral surveys that we can contrast with our daily social voting intentions (see Figure 3.4). Accordingly, we assume that electoral surveys are effectively able to catch the "true" evolution of an electoral campaign, only possibly with some time-gap. This time-delay should not be present in the social media monitoring case, however, given the real-time nature of such data.

When working with two (or more) time series, the usual need is to study their correlation but also potential causation effects: who leads who? This technique is called *lead-lag* estimation. The lead-lag effect is a concept of common practice that has some history in financial econometrics and that we bring to social media analysis. In time series, for instance this notion can be linked to the concept of Granger causality, and we refer to Comte and Renault (1996) for a general approach. From a phenomenological perspective, the lead-lag effect is supported by empirical evidence reported in Chiao et al. (2004), de Jong and Nijman (1997) and Kang et al. (2006), together with Robert and Rosenbaum (2011) and the references therein. The usual Granger-like approach has several limitations: 1) the time series must be of the same frequency (daily data, weekly date, etc.); 2) testing for causality often leads to discover a bidirectional effect; 3) linear time series are used (VAR or similar) to model the data. An additional problem is that, if the frequency of time series increases, that is the lag between the data diminishes, the empirical correlation vanishes artificially due to the so-called Epps effect (Zhang 2011).

In most applications, data usually have different frequencies (social media versus survey data), contain missing data (data are not always available) and hence are also asynchronous, and there is no reason to assume a linear behavior (in many cases we observe spikes, etc.). To take into account all the mentioned features of the data, Hoffmann et al. (2013) proposed a lead-lag estimator based on the

Hayashi–Yoshida asynchronous covariance estimator (Hayashi and Yoshida 2005, 2008). This estimator also overcame the Epps effect, so it works well in the presence of high frequency data.

Let $\theta \in (-\delta, \delta)$ be the time lag between the two nonlinear processes X and Y. Roughly speaking, the idea is to construct a contrast function $U_n(\theta)$ which evaluates the Hayashi–Yoshida estimator for the times series X_t and $Y_{t+\theta}$ and then to maximize it as a function of θ. The lead-lag estimator $\hat{\theta}_n$ of θ is defined as

$$\hat{\theta}_n = \arg \max_{-\delta<\theta<+\delta} |U_n(\theta)|.$$

When the value of $\hat{\theta}_n$ is positive it means that X_t and $Y_{t+\hat{\theta}_n}$(or $X_{t-\hat{\theta}_n}$ and Y_t) are strongly correlated, so we say "X leads Y by an amount of time $\hat{\theta}_n$", so X is the leader and Y is the lagger. Vice versa for negative $\hat{\theta}_n$.

Now, let's go back to the electoral surveys and to our social media analysis, and let's compare the two time series using the lead-lag method. In particular, let's focus on the difference between the Obama and Romney expected vote share according to the two previous estimations along the period illustrated in Figure 3.4 (i.e. from 28th of September till 5th of November 2012).

Table 3.5 provides the estimates of the lead-lag analysis while Figure 3.5 gives a graphical representation of the results. Positive values of the lag $\hat{\theta}_n$ would indicate that the variable in row anticipates the variable in column by a certain number of days. Conversely, negative values of the lag $\hat{\theta}_n$ would indicate that the variable in row is anticipated by that in column.

The results suggested that estimation of social media comes first, tending to anticipate, in the short term, the trend recorded by electoral surveys on average by 8 days. This comes as no surprise given the effort in sampling, data collection and data analysis of survey polls, that makes basically impossible a real-time analysis based on surveys. Having said that, our analysis illustrated that if survey polls, with their natural reporting lags, may not always be timely indicators of an underlying social phenomena (such as an electoral campaigning), the opposite happens with social media data – at least in the 2012 US case. In the light of this, such additional study has confirmed the accuracy of social media data not only in forecasting the final results of the elections but also in monitoring daily the evolution of voters' preferences and their reactions to all the events that characterize the unfolding of any electoral campaign.

Table 3.5 Estimates of the lead-lag analysis in the short term

	Social Voting Intentions (SASA)
Electoral surveys	−8* (0.028)

Note: Entries are number of days. P-values in parentheses. Significance (two tailed): * .05.

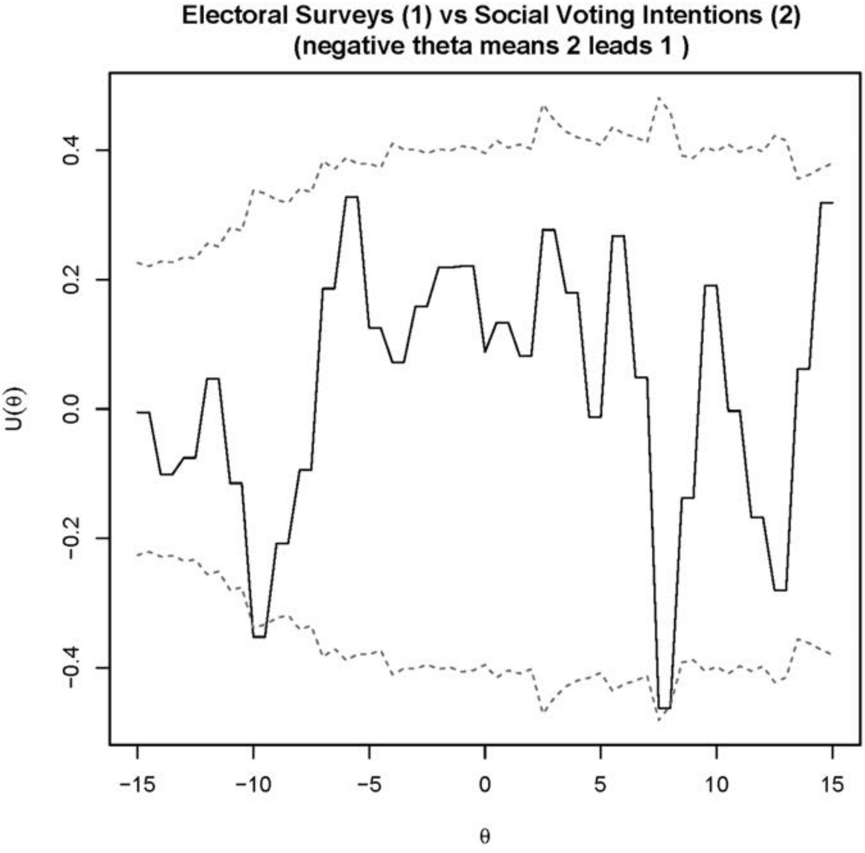

Figure 3.5 Estimates of the lead-lag analysis in the short term

Note: A value of $U(\theta)$ outside the bounds implies a significant lag. In this case, this happens when $\theta = 8$.

The selection of the centre-left coalition leader in Italy (2012)

We double-checked the predictive skills of social media analysis by applying the SASA technique to forecast the first and second rounds of the primary elections held by the centre-left coalition called “Italia Bene Comune”, to select the leader of the alliance. This is an intriguing exercise, given the particularly complex environment of a primary election that makes the possibility of a forecast particularly burdensome (AAPOR 2009): primary elections are, indeed, a contest that involves typically a partisan electorate, a larger number of viable candidates, with less ideological differentiation than one finds in the general election (a fact that makes more costly the electoral choices by voters) and lower turnout (Jensen and Anstead 2013). The fact that in Italy the primary elections are not legally recognized as such just makes things harder.

The centre-left coalition Italia Bene Comune was composed of four parties. The main one was the Democratic Party (PD). It formed an alliance with the left-wing Left Ecology and Freedom (SEL), the small Italian Socialist Party (PSI) and the Democratic Centre (CD). In autumn 2012, before the beginning of the campaign for the 2013 Italian general election, this coalition decided to select its leader through primary election. Such election was opened to all the voters that were members or sympathizers of one of the parties involved in the coalition. Five different candidates contested the election: Pier Luigi Bersani (leader of the PD), Matteo Renzi (Mayor of Florence, PD), Nichi Vendola (governor of Apulia and leader of SEL), Laura Puppato (party whip in Veneto, PD) and Bruno Tabacci (Milan budget councilor and leader of CD).

For such analysis, we adopted the SASA in a twofold manner. First, we decided to exploit the (online) wisdom of the crowd of social media (Franch 2013) to see whether online comments were able to accurately predict the winner of the election. Second, we measured the preferences expressed online to detect the voting intentions toward each candidate.

2012 Centre-Left primary election (1st round)

Period	6 October–25 November 2012 (with gaps)
Data Source	Twitter
Data Gathering	API
Keywords	bersani, renzi, vendola, puppato, tabacci, primariepd, lovotoperche, primarie2012, matteorenzi, pbersani, LauraPuppato, NichiVendola, TabacciPrimarie, oppurevendola, primarie pd, adesso renzi, tuttixbersani, pb2013, renzi2012
No. comments	495,003
SASA Method	iSA
Weights	No
Election day	25 November
Forecast release	Daily (from 21 November); last release: 25 November (9.00, CET)
Reference	http://www.corriere.it/politica/speciali/2012/primarie-centrosinistra/; https://voicesfromtheblogs.wordpress.com/2012/11/27/twitter-benecomune-previste-anche-le-primarie/

The first analysis was made on 6 October, before the campaign started and the candidacies became official. It has been argued that one way to predict the election is to examine the expectations of voters about which candidate will win, rather than looking at their preferences (Lewis-Beck and Skalaban 1989; Lewis-Beck and Stegmaier 2011; Rothschild and Wolfers 2011). We measured the expectations that emerged on online discussion about which candidate could win the primary. Although the prediction was made well in advance, the outcome was very accurate.

As early as the beginning of October, the majority of comments (55.5%) already foresaw a victory of Bersani; Renzi was ranking second, as 27.4% of comments predicted his victory, while only 17.1% speculated about the victory of Vendola,

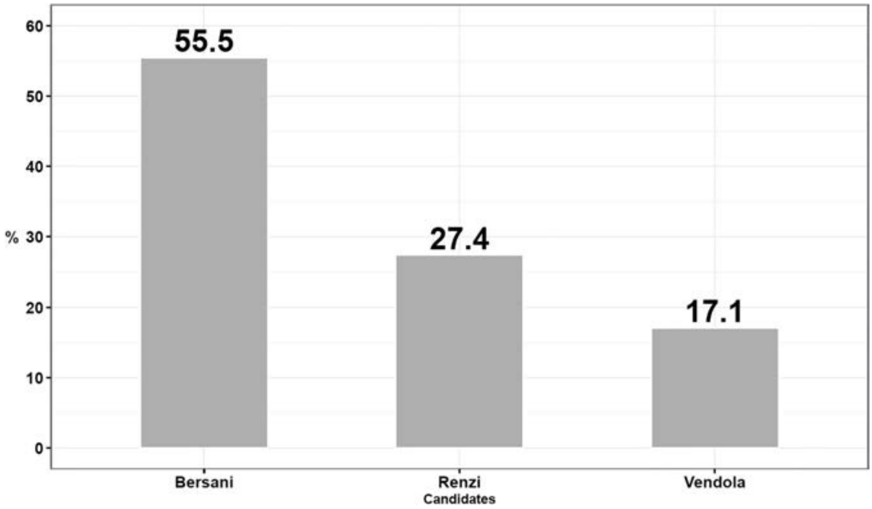

Figure 3.6 General expectations on Twitter about the outcome of the 2012 centre-left primary election (SASA)

and nobody discussed the chances of Puppato and Tabacci (see Figure 3.6). Indeed, almost two months later, Bersani actually won the second round of the primary election, becoming the leader of the centre-left. In addition, the ranking too was correctly predicted given that Bersani won the runoff against Renzi, while Vendola ranked third in the first ballot. Finally, assuming that these shares (indicating the probability that each candidate could win) could be matched with the actual votes gained by these candidates, we would observe huge similarities. The MAE of such prediction would be respectively 4.8 (in the first ballot) and 6 (in the second ballot), that was in line with those of many surveys. This can be taken as a further signal that Twitter users were really able to discriminate the strength and the electoral appeal of each candidate. As such, even when we are temporally distant from an election, the information available online can become a precious source to guess which candidate can be the best one, being more appealing to voters and retaining the highest chances to get elected.

Beside the expectations, we also measured the preferences. From 6 October to 25 November (the day the election was held) we analyzed around 500,000 tweets at several points in time to assess the voting intention of Twitter users toward these five candidates. Figure 3.7 reports the fluctuation of voting intentions, according to our analysis. We reported our estimates measured in 10 different days by analyzing around 40,000–50,000 tweets released in a time span that ranged from 10 days (at the beginning of the electoral campaign) to 1 week (since 12 November).

Following what we have done in the French and US analyses, Figure 3.7 allows once again to monitor how the distribution of preferences changed over

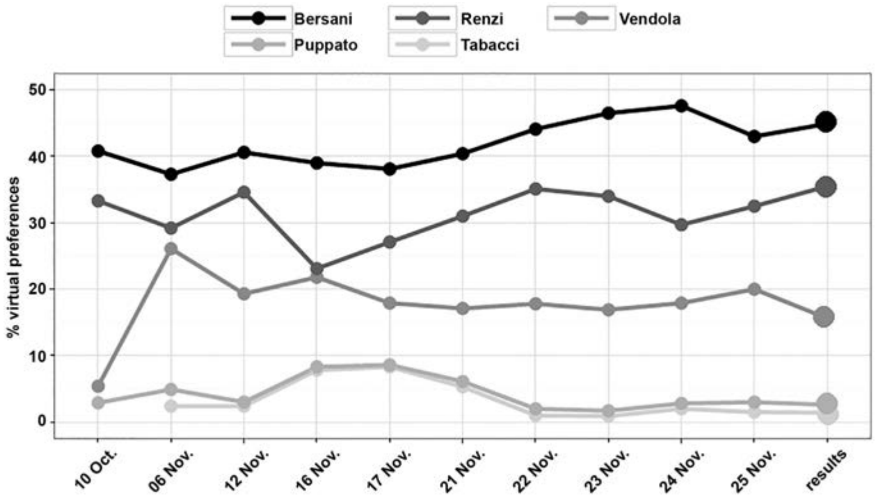

Figure 3.7 Centre-left primary election, first round: candidates' share of votes according to SASA and comparison with actual results

the campaign as well as the different "momentums" that characterized such campaign. To start with, the expected votes share of Bersani was always around 40% while we observed an increase during the last week, when the voting intentions grew to 47.6% (note that a similar trend was reported also by several polls). However, in the last two days before the elections, this value shrank to 43%, closer to the actual result (we were mistaken by 1.9% only). Renzi always ranked second since the beginning. His support was on average around 31%, with a peak on 12 November, when the candidates were involved in a debate on SKY television, followed by a loss after the debate.

The loss in the voting intentions expressed toward Renzi could have been due to his performance in the debate. Indeed, the debate represented the first original attempt to live broadcast an episode of struggle between candidates belonging to the same coalition (and some to the same party too). It markedly attracted the attention of citizens. Almost 2 million people (1,885,816) watched it on television, with a share of 6.22%. In the 2 hours of debate, up to 20 tweets per second were written to comment on it, and the total number of tweets approached 100,000. By analyzing a subsample of these tweets, we noticed that a plurality of comments (34%) claimed that Bersani won the debate.[23]The performances of Renzi (32.8%) and Vendola (26.4%) were appreciated too. To summarize the debate in a single tweet we can observe that voters "*would buy a second hand car from Bersani, a book from Vendola and a dream from Renzi*".

Although the debate was balanced, both Renzi and Vendola were expected to perform markedly better than the front-runner Bersani, and they failed to fulfill such expectations. This was particularly true for Renzi, as many analysts claimed that he could have exploited the debate to close the gap with Bersani. To the contrary, the PD leader proved able to survive to the debate and his approval has not been challenged by the performance of Renzi. Accordingly, after the debate the enthusiasm of Renzi's supporters dampened, and his level of support decreased. Renzi's expected votes share started to grow again only after the convention (called "Leopolda 2012") held in Florence and organized by Matteo Renzi and his supporters. Nichi Vendola, who actually ranked third, started his campaign only in November after he was discharged from a prosecution.[24] This combination of events attracted him an initial large share of support, which declined in a few days when such effect vanished, bringing Vendola's expected votes share around 18%. Finally, the two minor candidates, Puppato and Tabacci, retained only a low share of votes (according to our forecast) during the whole campaign, except after the debate when they took advantage from an outstanding public visibility.

The similarities between our forecasts (which were published daily on the online version of the main Italian newspaper, *Corriere della Sera*) and the final outcome of the election were once again pretty huge. This is true both with respect to our final prediction, released on 24 November, and for an "experiment" based on a kind of "exit poll" made on 25 November by analyzing tweets published on Election Day.[25]

Figure 3.8 also reports the actual votes share and highlights, per each candidate, the absolute difference between our last forecast and the final results. The gap between our estimates and the results was very narrow, being on average below 2%. This error was in line with the average error provided by the surveys

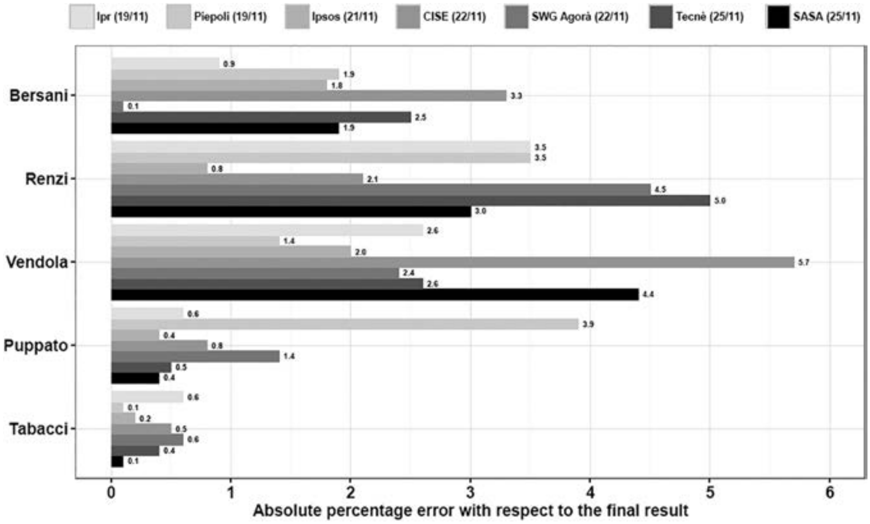

Figure 3.8 First round of the 2012 Italian centre-left primary elections: absolute difference between final results and predictions according to survey polls and SASA

Table 3.6 Comparison of the accuracy of Twitter forecast and surveys polls in the Italian primary election of the centre-left coalition (first round)

	Gap Bersani–Renzi
Popular Vote	*Bersani +9.4*
SASA	*Bersani +10.5*
Piepoli	Bersani +11.0
Ipr	Bersani +5.0
SWG	Bersani +14.0
CISE	Bersani +10.6
Tecnè	Bersani +16.9
Ipsos	Bersani +8.42

polls issued in that last week, which was 1.9%. This also appears from Table 3.4 where we display, per each candidate, the absolute error of our prediction along with those of the polls. What is more, our technique succeeded in predicting the gap between the two foremost candidates, Bersani and Renzi, better than the traditional survey polls. Indeed, according to our results, the gap between the two candidates was 10.5% while Bersani led by 9.4 points after votes were counted (Table 3.6). This means a difference of 1.1% while on average the polls mistook the magnitude of the gap by 3%.

We were able to produce a similarly accurate forecast of the Italian centre-left primaries also in the second round, contested by Bersani and Renzi and held on 2 December. The prediction was made in the evening of 1 December and was publicly issued during a show (*#votantonio*) broadcast by Radio24.[26]

2012 Centre-Left primary election (2nd round)	
Period	29 November–1 December 2012
Data Source	Twitter
Data Gathering	API
Keywords	bersani, renzi, lovotoperche, voto renzi, voto bersani, adesso renzi, pb2013, tuttixbersani, pbersani, matteorenzi, oppurerenzi, votomatteo, votero renzi, votero bersani, ballottaggio, primarie csx, bersanipresidente, votorenzi, votobersani
No. comments	24,783
SASA Method	iSA
Weights	No
Election day	2 December
Forecast release	1 December (19.00, CET)
Reference	Radio24, radio show: #votantonio; https://voicesfromtheblogs.wordpress.com/2012/12/01/twitter-primarie-ore-19-bersani-58–4-renzi-41–6/

Table 3.7 Comparison of the accuracy of Twitter forecast and surveys polls in the Italian primary election of the centre-left coalition (second round)

	Day of publication of the survey	*Bersani*	*Renzi*	*Gap*
Popular Vote	–	*60.9*	*39.1*	*Bersani +21.8*
SASA	*1/12/2012*	*58.4*	*41.6*	*Bersani +16.8*
Piepoli	25/11/2012	59.0	41.0	Bersani +18.0
Ips	26/11/2012	56.0	44.0	Bersani +12.0
ISPO	27/11/2012	56.5	43.5	Bersani +13.0
SWG	28/11/2012	55.0	45.0	Bersani +10.0
COESIS	28/11/2012	54.0	46.0	Bersani +8.0
Quorum	28/11/2012	56.4	43.6	Bersani +12.8
Ipsos	29/11/2012	57.5	42.5	Bersani +15.0

Table 3.8 Comparison of the accuracy of Twitter forecast made through mentions, automated sentiment analysis and SASA method (Italian primary election of the centre-left coalition, first round)

Method	*Source*	*MAE*
Mentions	http://seigradi.corriere.it/2012/11/25/le-primarie-del-centrosinistra-su-twitter-vincono-renzi-e-vendola/	6.36
Mentions	http://www.chefuturo.it/2012/11/twitter-la-tv-e-i-voti-reali-analisi-del-primo-round-delle-social-primarie/	9.72
Sentiment Analysis	http://vincos.it/2012/11/25/primarie-centro-sinistra-citazioni-e-performance-online-dei-candidati/	8.65
SASA	–	*1.96*

Table 3.7 reports the actual results, our forecast (made by using Twitter data) and the results of several surveys published in the last week before the second round. In this case, our analysis on almost 25,000 tweets posted between Thursday 29 November and Saturday 1 December (the night ahead of the second round) predicted a clear victory for Bersani (58.4% versus 41.6% for Renzi). At the ballot, Bersani won with 60.9% of votes against 39.1% for Renzi.

According to these results, social network sites confirm themselves as sources of valuable information that can be exploited to carry out electoral forecasts. However, the goodness of such forecasts seems once again to depend on the technique adopted. Table 3.8 portrays a comparison of the MAE obtained by the SASA method versus the ones arising from several other social media analyses conducted on the first round of the Italian primary elections that considered either the volume of mentions or the positive sentiment measured through dictionary-based methods.[27] As can be seen, the MAE increased a lot in the latter types of analyses. What is worse, such analyses clearly (and consistently) highlighted a strong advantage for the actual second-ranked candidate, Matteo Renzi.[28]

"Win some, lose some"? Twitter and the selection of party leaders

After the successful prediction of the centre-left primary election, we attempted to predict two other similar elections related to the selection of party leaders. In December 2013, the Democratic Party (PD) and the Northern League (LN) – two of the main Italian parties – selected their leaders through direct election (so-called primaries). On 7 December, the LN organized "closed primaries" in which only party members were allowed to vote. Conversely, on 8 December, the PD selected the leader through "open primaries", in which anyone could vote after declaring his or her affinity with the party.

2013 Northern League 'primary' election

Period	22 November–6 December 2012
Data Source	Twitter
Data Gathering	API
Keywords	salvini, bossi, leganord, congresso lega, segretario lega, primarie lega, firme lega, segreteria lega, voto lega, voto bossi, voto salvini, votato bossi, votato salvini, voterò salvini, voterò bossi, firmato bossi, firmato salvini, iovotosalvini, leganord2, primarielega
No. comments	8,096
SASA Method	iSA
Weights	No
Election day	7 December
Forecast release	6 December (evening)
Reference	http://voicesfromtheblogs.com/2013/12/06/non-ce-solo-il-pd-anche-la-lega-ha-le-sue-primarie-e-le-vincera-salvini-almeno-su-twitter/

Only two candidates competed to become the new LN leader. The front-runner Matteo Salvini was challenged by the founder of the LN and former party leader, Umberto Bossi. The electorate was composed of very few people, as only 17,000 party members were entitled to vote. Only 10,221 actually cast a vote for one of the two candidates. On the day before the election, we published our prediction based on SASA analysis of more than 8,000 tweets that discussed the selection of the LN leader and that were published in the last 2 weeks before the election.[29] Such number was remarkably higher if compared to the actual number of voters. We predicted the victory of Salvini with 82.5% of votes, whereas he actually won 82% of support. Hence, our prediction was very accurate, leading to a low MAE (0.5).

2013 Democratic Party 'primary' election

Period	18 November–8 December 2013
Data Source	Twitter
Data Gathering	Firehose data provider
Keywords	renzi, cambiaverso, cuperlo,belloedemocratico, civati, civoti, pd, primariepd, congressopd, segretariopd, 8dicembre, primarie pd, segretario pd, congresso pd, civatisegretario, iostoconcivati, cuperlosegretario, giannicuperlo, matteorenzi, nientetrucchi, quicuperlo, belliedemocratici, renzisegretario, quirenzi, voto cuperlo, voto civati, voto renzi, voto primarie, scelgo cuperlo, scelgo civati, scelgo renzi, adessopartecipo, comitatocuperlo, convenzionepd, convenzione pd, voterò renzi, voterò cuperlo, voterò civati, votato renzi, votato cuperlo, votato civati, iovotoperchè, iovotoperché, occupypd, ilconfrontopd, confrontopd, pdsky, skypd, pdfactor, iovotomatteo, votocuperlo, anchebasta, megliocuperlo, lovotoperché, lovotoperchè, votomatteo, unaltramusica, ungiornonuovo, sevincerenzi, sefossirenzi, preferenzi, vincecivati, voltabuona, lavoltabuona, matteorisponde
No. comments	607,710
SASA Method	iSA
Weights	Yes: (1) party congress votes; (2) expected turnout; (3) online activists
Election day	8 December
Forecast release	Daily; last release: 8 December (8.00, CET)
Reference	http://www.corriere.it/politica/speciali/2013/primarie-pd/

The selection of the PD leader was far more complex. It was based on a two-stage process. In the first stage, only party members were entitled to vote. Between 7 November and 17 November, local party congresses selected the local party leaders and at the same time voted for their preferred national party leader. Four candidates were competing during this stage: Matteo Renzi, Gianni Cuperlo, Giuseppe "Pippo" Civati and Gianni Pittella. Almost 300,000 party members voted, and Renzi ranked first with 45.3% of votes. Cuperlo was second (39.4%), while the votes share of Civati (9.4%) and Pittella (5.8%) was much lower.

In the second stage, the selection of the leader was opened also to PD sympathizers who were not party members. A wider (and undefined) number of citizens was then allowed to cast a vote. As party rules allow only three candidates to run in open primaries, the fourth candidate, Pittella, was excluded and he decided to endorse Renzi. The actual campaign for the leadership started on 18 November, after the end of the PD congress.

From that day until 8 December we analyzed more than 600,000 tweets to produce a daily forecast. The estimates of the voting intentions expressed toward the three candidates were published every day on the online version of the main Italian newspaper, *Corriere della Sera*.[30]

While in the previous forecasts presented in this chapter we did not apply any sort of weighting, here we decided to experiment with a different (and rather complex) strategy. In fact, while it is true that the population of potential voters of "PD primaries" was undefined, as long as the open primaries were organized after the PD congress, we could hypothesize (as we actually did) that all party members who voted during the congress would have voted again during the primaries. We assumed that their vote choice would not have changed. Accordingly, we produced our Twitter forecast taking this information into account. We also weighted the results as follows.

By using the number of comments discussing the PD primaries (in contrast with the number of comments that discussed about the IBC centre-left primaries in 2012) as a proxy for the expected turnout, we estimated that around 1.5–2 million voters would have cast a vote on 8 December. Arguing that, among those 1.5–2 million, the expected vote of 300,000 voters (PD members who voted during the congress) was already known, we produced a weighted forecast in which the Twitter prediction (weighted by 1.2–1.7 million) was balanced by the actual results of party congress (weighted by 0.3 million).

Figure 3.9 reports the daily variation in voting intentions expressed toward the three candidates. We noticed that Renzi was always leading by a wide margin. His votes share was always around 50%, with a marked increase in the last days before the election when Renzi actually mobilized his activists to organize campaign stands in all the main Italian cities. The final slogan launched by Renzi in the last days of campaign (*#lavoltabuona*, i.e. "this is it") also proved to be very incisive in increasing his voting intentions. Conversely, the support toward Cuperlo and Civati (the two candidates belonging to the left-wing factions of the party) was always lower. What is more, we observed that the increase in the voting intentions toward one candidate was usually reached at the expense of the other. The

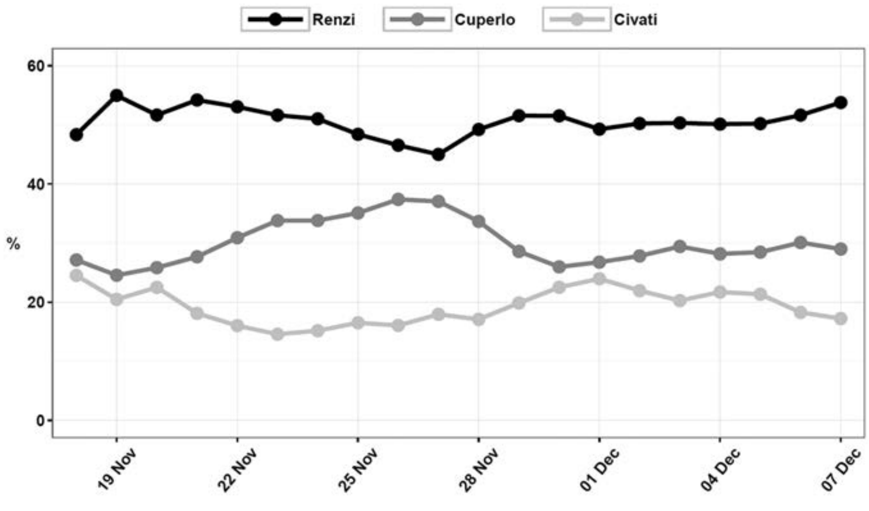

Figure 3.9 Italian Democratic Party primary election: daily social voting intentions according to SASA

support for Civati decreased a lot after his choice to toe the party line in a key parliamentary vote, when he decided not to support a "motion of no confidence" proposed against one minister of the Letta (PD) government, Annamaria Cancellieri, who was accused of abuse of power by the opposition. Another key moment of the campaign was represented by the television debate held on 29 November. On that day, a huge number of comments (76,000) was posted online. Both Civati and Renzi took advantage of the debate, showing their rhetoric ability, and their online support grew accordingly.[31] To the contrary, the performance of Cuperlo during the debate was very poor, and this definitively washed out any possibility that Cuperlo could challenge Renzi in the race for the leadership.

According to this analysis, the final prediction was issued on 8 December, early morning, when the polls opened.[32] We correctly predicted the victory of Matteo Renzi as well as the ranking of the three candidates. However, the average error (9.17) was quite large if compared to the previous predictions presented in this chapter. Remarkably, the error was also affected by the choice of weighting the Twitter voting intentions. In fact, the MAE would have been 30% lower if we had stuck to the usual non-weighted prediction, leading to an average error of 6.3 (see Table 3.9). This value is still larger than usual; however, we should acknowledge that in this election the average error of surveys was also higher if compared to other elections. The mean value of MAE for the surveys held in the last weeks before the election was in fact 4.03, which was relatively close to the MAE of our non-weighted Twitter prediction. In Chapter 5 we provide a more detailed discussion of the factors that can increase (or decrease) the accuracy of social media electoral predictions.

Nevertheless, this case is interesting for a number of reasons. First, despite anecdotal reports made by journalists that suggested on Twitter Civati was ranking first,[33] our analysis showed that this was not the case. To the contrary, Renzi was consistently ahead in the preferences declared on social media (for a similar result obtained through machine learning techniques, see Coletto et al. 2015). This further emphasized the need to investigate social media by using proper methods of analysis.

Second, this case study also provided the opportunity to analyze the relationship between the Twitter voting intentions and the degree of negativity expressed online. In fact, beside the voting intentions, we also measured the daily degree of

Table 3.9 Italian Democratic Party primary election: contrasting the actual votes of the candidates with Twitter (SASA) predictions (both weighted and non-weighted)

	Actual votes	*Weighted Twitter prediction*	*Non-weighted Twitter prediction*
Renzi	67.55	53.8	58.1
Cuperlo	18.21	29.0	21.1
Civati	14.24	17.2	20.8
MAE		9.17	6.30

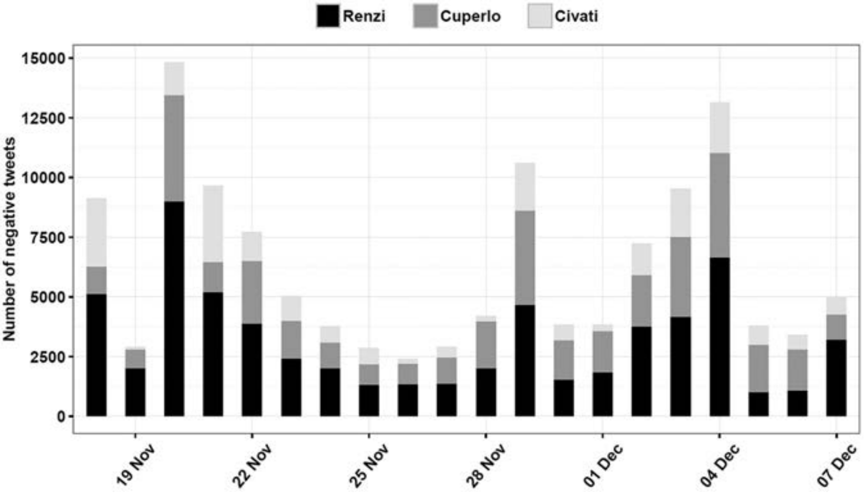

Figure 3.10 Italian Democratic Party primary election: daily evolution of the level of negative campaigning in Twitter discussion over candidates

negativity, that is the amount of negative comments expressed in online conversations and the relative share of negative sentiment of each candidate.[34]

Figure 3.10 reports the daily evolution of the level of negativity. On average, 22% of comments published online expressed criticism toward one of the candidates. This rate is in line with what was observed in other elections. For instance, Hosch-Dayican et al. (2014) noticed that in the 2012 Dutch parliamentary campaign, on average, 20% of comments released on Twitter by citizens who tried to persuade others was composed of "negative campaigning" messages. This share is also slightly higher than average share of negative campaign messages written by official accounts of parties and candidates, which is around 10%–15% (e.g. Ceron and Curini 2016; Ceron and d'Adda 2015; Evans et al. 2014; Hosch-Dayican et al. 2014).

The share of negativity was markedly higher in the first week of campaign, when almost 33% of tweets expressed negative campaigning. In the following days the rate of negativity decreased up to the point that, in the last three days before the election, only 10% of comments were negative. Across the campaign we noticed three peaks of negativity. The first one, as we said, was located at the beginning of the campaign. This was caused by two different events. On the one hand, the front-runner candidate, Matteo Renzi, was indirectly accused of electoral frauds in some local party congresses. Immediately after the end of the party congress (from 18 November) many users attacked him in relation to such episodes. Around 60% of negative messages written in those days contained attacks to Renzi. Conversely, only a few days later another (already mentioned) episode entered into the agenda of the campaign. Civati's refusal to vote a motion of no confidence proposed by the opposition against the Minister Cancellieri generated a heated debate online and produced a strong increase in the share of negative messages written against Civati. The day before the television debate was also characterized by a big level of negativity; while the three candidates did not made large use of negative campaigning on television, their supporters and their activists were much less polite in Twitter conversations. Finally, the third peak appeared on 3–4 December, when Renzi attacked the government, raising the tone of the campaign. However, in the last days before the election more attention was devoted to positive messages in the attempt to support one's own favorite candidate and to broadcast the image of a unified party.

What is particularly interesting is the fact that the share of negative messages is unrelated to the voting intentions expressed online. For all the three candidates, in fact, such relationship is not statistically significant. Indeed, Renzi was the most criticized candidate every day (except 5 December) and he attracted 50% of the criticism (32% of comments were written to criticize Cuperlo and 18% to criticize Civati). This is not surprising if we consider that here we are talking about an intra-party election, and in elections like such (by definition) party activists, members and sympathizers are involved. Moreover, negative comments can also be considered as a typical example of *negative campaigning*. In this respect, the large literature base on negative campaigning reminds us that the front-runner candidate (in this case, Matteo Renzi) usually attracts more negative campaign messages (Skaperdas and Grofman 1995).

But even if not all negative messages are written to perform negative campaigning, there are still doubts in considering them when making electoral forecasts based on social media data, given that those comments are not really expressing any voting intention for a party/candidate. In a two-party system, where only two candidates are contesting the race, we still cannot really judge – from a negative comment – whether the voter is really willing to support the opponent of the candidate the he or she criticized rather than choosing to abstain (see our previous discussion on the rules applied to map tweets into votes). In a multiparty scenario, our inability in connecting one negative post to a positive vote for a party/candidate just gets worse.

For analogous reasons, as just noted, the idea of subtracting negative comments from positive ones when evaluating the online strength of different candidates is also risky (Jungherr et al. 2016). Precisely because the front-runner is more likely to attract negative comments, and given that a mere negative sentiment is meaningless in term of voting intentions as just noted, such operationalization would lead to biased predictions.

In the next chapter, we dig more into positive and negative messages by analyzing the campaign effects of several features that can affect voters' choice such as positive campaigning (particularly, policy promises) and negative campaigning. We also take into account the impact of the online valence endowment of political leaders. By doing that, we show how social media data can also be exploited to catch the dynamics of electoral campaigns and to evaluate the efficacy of different campaign strategies also providing interesting new theoretical political insights.

Notes

[1] The rate of penetration is defined as the number of monthly active Twitter users relative to the number of Internet users.

[2] Peerreach.com (http://blog.peerreach.com/2013/11/4-ways-how-twitter-can-keep-growing/)

[3] Notice that when the period of analysis was sufficiently long and the daily amount of data was sufficiently large we applied a 3- or 7-day moving average that allowed us to take into account multiple daily forecasts and capture the trend which is a nowcast feature of social media data. According to our experience, a 7-day moving average is usually the preferable choice (see also O'Connor et al. 2010) provided that enough points in time are available in the analysis (when the time series was shorter, however, we reverted to a 3-day moving average). We also tried different methods such as the dynamic linear model without obtaining large difference in the final estimations (Walther 2015).

[4] The proportion of geo-referenced data amounts to 5% to 10% of the entire volume of tweets.

[5] In doing that, all duplicate texts were always removed.

[6] Of course, applying such rules makes (much) more sense within a SASA approach rather than under a completely automated approach.

[7] See: http://voicesfromtheblogs.com/2012/04/22/presidenziali-francesi-su-twitter/

[8] Source: http://www.vanksen.fr/blog/e-presidentielle-hollande-toujours-derriere-sarkozy/

[9] Note that both in this and in the following analyses, some keywords were added across the campaign to take into account the evolution of the language over time, in order to try to reach the whole population of comments and to take into account that new hashtags or now keywords could have been used by politicians, parties and voters. However, our results were substantially the same also when analyzing – thorough the campaign – a fixed set of keywords composed of a balanced number of terms (i.e. *hollande*, *sarkozy*, *presidentielle*, *election*, *elysee*, *Objectif2012*, *RadioLondres*, *UMP election*, *PS election*, *FH2012*, *NS2012*).

[10] Applying a shorter moving average (i.e. 3 days) did not change qualitatively any of our results. See also note 3 in this chapter.

[11] Source: http://www.corriere.it/esteri/speciali/2012/elezioni-usa/

[12] Notice that other comments were merely ironic without expressing any opinions. For instance "*Romney wins the debate on hair alone*" or "*Romney has great hair*." Thanks to SASA, these comments were not considered in the analysis. Conversely, an automated tool of sentiment analysis would have probably considered these as positive messages given that they include "positive" words like *wins* or *great*.

[13] This ratio was measured by considering only comments suggesting that one of the two was the winner of the debate. Source: http://edition.cnn.com/2012/10/03/politics/debate-main/

[14] Source: http://edition.cnn.com/2012/10/16/politics/debate-mainbar/

[15] Source: http://edition.cnn.com/2012/10/22/politics/debate-mainbar/

[16] First debate: http://sentimeter.corriere.it/2012/10/04/romney-obama-per-la-rete-e-cappotto/; second debate: http://sentimeter.corriere.it/2012/10/17/visti-dalla-rete-stavolta-il-dibattito-premia-obama/; third debate: http://sentimeter.corriere.it/2012/10/23/per-la-rete-obama-vince-ancora-ma-bastera/

[17] To compare sentiment analysis and surveys data properly, only comments and answers that clearly indicate Obama or Romney as the winner of the debate were considered.

[18] See for instance the article by Andrew Gelman, director of the Applied Statistics Center at Columbia University: http://campaignstops.blogs.nytimes.com/2012/10/30/what-too-close-to-call-really-means/?smid=tw-share

[19] As said, the percentage of geo-referenced tweets is usually a small portion of the overall number of tweets posted every day. As such, the sample we drew upon was necessarily

global in nature, while not being necessarily representative of those using Twitter. Geo-referenced information was taken from the GPS coordinates contained in the metadata of the tweets or through the Twitter API directly which allowed us to collect data in a prescribed vicinity of a given city/location using self-declared users' information.

[20] Twindex, a fully automated index based on a dictionary-based method, was constructed for the two candidates as the proportion of the total number of positive tweets (measured by means of ontological dictionaries) versus the total number of positive *and* negative tweets. See http://usatoday30.usatoday.com/news/politics/twitter-election-meter

[21] Washington et al. (2013) focused in their analysis on the proportion of positive sentiment messages only.

[22] Choy et al. (2012) measured support by considering both positive and negative sentiment.

[23] Source: http://sentimeter.corriere.it/2012/11/13/csx-factor-la-rete-si-infiamma-ma-tra-bersani-e-renzi-e-pareggio/

[24] Note that on 10 October the campaign for the primary elections was not officially started yet. Therefore, our estimates included some potential candidates who later decided to not run for nomination.

[25] Sources: http://www.corriere.it/politica/speciali/2012/primarie-centrosinistra/and http://sentimeter.corriere.it/2012/11/27/anche-per-le-primarie-twitter-ci-azzecca/

[26] Source: http://voicesfromtheblogs.com/2012/12/01/twitter-primarie-ore-19-bersani-58–4-renzi-41–6/

[27] Unfortunately, to our knowledge, no social media analysis beside ours on the second round of the Italian primary elections has been conducted.

[28] Sources for the analyses reported in Table 3.8 are as following: Mentions 1 (http://seigradi.corriere.it/2012/11/25/le-primarie-del-centrosinistra-su-twitter-vincono-renzi-e-vendola/); Mentions 2 (http://www.chefuturo.it/2012/11/twitter-la-tv-e-i-voti-reali-analisi-del-primo-round-delle-social-primarie/); automated sentiment analysis (http://vincos.it/2012/11/25/primarie-centro-sinistra-citazioni-e-performance-online-dei-candidati/).

[29] See: http://voicesfromtheblogs.com/2013/12/06/non-ce-solo-il-pd-anche-la-lega-ha-le-sue-primarie-e-le-vincera-salvini-almeno-su-twitter/

[30] See: http://www.corriere.it/politica/speciali/2013/primarie-pd/

[31] Indeed, according to survey polls, these two candidates were judged as the winners of the debate (each candidate gathered the approval of 37% of the respondents, while only 18% thought Cuperlo had won). See: http://www.gadlerner.it/2013/11/30/confronto-pd-per-sondaggi-e-twitter-vince-civati

[32] See: http://sentimeter.corriere.it/2013/12/08/primarie-pd-e-per-la-rete-the-winner-is/

[33] See: http://www.corriere.it/politica/speciali/2013/primarie-pd/notizie/primarie-pd-twitter-vincecivati-29ea5b52–5c02–11e3-bc7d-29ea5b52–5c02–11e3-bc7d-68ebf7f6255f.shtml

[34] Source: http://sentimeter.corriere.it/2013/12/06/fratelli-coltelli-su-twitter-elettori-pd-ai-ferri-corti/

References

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013) 'Now-casting and the real-time dataflow'. Available at: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1564.pdf?74129250f82f3a27d5c798fc79f10dbb

Banks, J.S. (1991) 'Signaling games in political science', New York: Routledge.

Barclay, F.P., Pichandy, C., Venkat, A., and Sudhakaran, S. (2015) 'India 2014: Facebook "like" as a predictor of election outcomes', *Asian Journal of Political Science*, 23(2): 134–160.

Callander, S., and Wilkie, S. (2007) 'Lies, damned lies, and political campaigns', *Games and Economic Behavior*, 60(2): 262–286.

Cameron, M.P., Barrett, P., and Stewardson, B. (2013) ‘Can social media predict election results? Evidence from New Zealand’, Working Paper in Economics 13/08, University of Waikato.

Ceron, A., and Curini, L. (2016) ‘E-campaigning in the 2014 European elections: The emphasis on valence issues in a two-dimensional multiparty system’, *Party Politics*. Advance online publication. doi: 12:1354068816642807

Ceron, A., and d’Adda, G. (2015) ‘E-campaigning on Twitter: The effectiveness of distributive promises and negative campaign in the 2013 Italian election’, *New Media & Society*, 18(9): 1935–1955. doi: 1461444815571915

Chiao, C., Hung, K., and Lee, C. (2004) ‘The price adjustment and lead-lag relations between stock returns: Microstructure evidence from the Taiwan stock market’, *Empirical Finance*, 11: 709–731.

Choy, M., Cheong, M., Ma Nang, L., and Koo Ping, S. (2012) ‘US presidential election 2012 prediction using census corrected Twitter model’. Available at: http://arxiv.org/ftp/arxiv/papers/1211/1211.0938.pdf

Cogburn, D.L., and Espinoza-Vasquez, F.K. (2011) ‘From networked nominee to networked nation: Examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign’, *Journal of Political Marketing*, 10(1–2): 189–213.

Coletto, M., Lucchese, C., Orlando, S., and Perego, R. (2015) ‘Electoral predictions with Twitter: A machine-learning approach’. *IIR ’15: 6th Italian Information Retrieval Workshop 2015*, Cagliari, Italy, 25–26 May 2015.

Comte, F., and Renault, E. (1996) ‘Non-causality in continuous time models’, *Econometric Theory*, 12: 215–256.

Contractor, D., Chawda, B., Mehta, S., Subramaniam, L.V., and Faruquie, T.A. (2015) ‘Tracking political elections on social media: Applications and experience’, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 25–31 July 2015.

Davis, M.L., and Ferrantino, M. (1996) ‘Towards a positive theory of political rhetoric: Why do politicians lie?’, *Public Choice*, 88(1–2): 1–3.

De Jong, F., and Nijman, T. (1997) ‘High frequency analysis of lead-lag relationships between financial markets’, *Journal of Empirical Finance*, 4: 259–277.

Diaz, F., Gamon, M., Hofman, J.M., Kıcıman, E., and Rothschild, D. (2016) ‘Online and social media data as an imperfect continuous panel survey’, *PloS One*, 11(1): e0145406. doi: 10.1371/journal.pone.0145406

DiGrazia, J., McKelvey, K., Bollen, J., and Rojas, F. (2013) ‘More tweets, more votes: Social media as a quantitative indicator of political behavior’, *PloS One*, 8(11): e79449. doi: 10.1371/journal.pone.0079449

Evans, H.K., Cordova, V., and Sipole, S. (2014) ‘Twitter style: An analysis of how house candidates used Twitter in their 2012 campaigns’, *Political Science & Politics*, 47(2): 454–462.

Gayo-Avello, D. (2011) ‘Don’t turn social media into another “literary digest” poll’, *Communications of the ACM*, 54(10): 121–128.

Gordon, J. (2013) ‘Comparative geospatial analysis of Twitter sentiment data during the 2008 and 2012 U.S. Presidential Elections’, Master Thesis, University of Oregon.

Hayashi, T., and Yoshida, N. (2005) ‘On covariance estimation of non-synchronously observed diffusion processes’, *Bernoulli*, 11: 359–379.

Hayashi, T., and Yoshida, N. (2008) ‘Asymptotic normality of a covariance estimator for nonsynchronously observed diffusion processes’, *Annals of the Institute of Statistical Mathematics*, 60: 367–406.

Hoffmann, M., Rosenbaum, M., and Yoshida, N. (2013) 'Estimation of the lead-lag parameter from non-synchronous data', *Bernoulli*, 19(2): 426–461.

Hosch-Dayican, B., Amrit, C., Aarts, K., and Dassen, A. (2014) 'How do online citizens persuade fellow voters? Using Twitter during the 2012 Dutch parliamentary election campaign', *Social Science Computer Review*, 34(2): 135–152. doi: 0894439314558200

Huberty, M. (2015) 'Can we vote with our tweet? On the perennial difficulty of election forecasting with social media', *International Journal of Forecasting*. Available at: http://dx.doi.org/10.1016/j.ijforecast.2014.08.005

Jensen, M.J., and Anstead, N. (2013) 'Psephological investigations: Tweets, votes, and unknown unknowns in the republican nomination process', *Policy & Internet*, 5(2): 161–182.

Jungherr, A., Schoen, H., Posegga, O., and Jürgens, P. (2016) 'Digital trace data in the study of public opinion an indicator of attention toward politics rather than political support', *Social Science Computer Review*. Advance online publication. doi: 10.1177/0894439316631043

Kang, J., Lee, C., and Lee, S. (2006) 'Empirical investigation of the lead-lag relations of returns and volatilities among the kospi200 spot, futures and options markets and their explanations', *Journal of Emerging Market Finance*, 5: 235–261.

Khan, F.H., Bashir, S., and Qamar, U. (2014) 'TOM: Twitter opinion mining framework using hybrid classification scheme', *Decision Support Systems*, 57: 245–257.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013) 'Ontology-based sentiment analysis of Twitter posts', *Expert Systems with Applications*, 40(10): 4065–4074.

Lampos, V., and Cristianini, N. (2012) 'Nowcasting events from the social web with statistical learning', *ACM TIST*, 3(4), article number 72.

Lansdall-Welfare, T., Lampos, V., and Cristianini, N. (2012) 'Nowcasting the mood of the nation', *Significance*, 9(4): 26–28.

Lewis-Beck, M.S., Nadeau, R., and Bélanger, E. (2011) 'Nowcasting v. polling: The 2010 UK election trials', *Electoral Studies*, 30(2): 284–287.

Lewis-Beck, M.S., and Skalaban, A. (1989) 'Citizen forecasting: Can voters see into the future?', *British Journal of Political Science*, 19(1): 146–153.

Lewis-Beck, M.S., and Stegmaier, M. (2011) 'Citizen forecasting: Can UK voters see the future?', *Electoral Studies*, 30(2): 264–268.

MacWilliams, M.C. (2015) 'Forecasting congressional elections using Facebook data', *PS: Political Science & Politics*, 48(4): 579–583.

Mejova, Y., Srinivasan, P., and Boynton, B. (2013) 'GOP primary season on Twitter: "Popular" political sentiment in social media', *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining 2013*, Rome, Italy, 4–8 February 2013.

Mitchell, A., and Hitlin, P. (2013) 'Twitter reaction to events often at odds with overall public opinion', Available at: http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K.M. (2013) 'Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose', *Proceedings the Seventh International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, 8–11 July 2013.

Murthy, D. (2015) 'Twitter and elections: Are tweets, predictive, reactive, or a form of buzz?', *Information, Communication & Society*, 18(7): 816–831. doi: 10.1080/1369118X.2015.1006659

Nooralahzadeh, F., Arunachalam, V., and Chiru, C. (2013) 'Presidential elections on Twitter–An analysis of how the US and French election were reflected in tweets', Paper presented at the 19th International Conference on Control Systems and Computer Science, Bucharest, Romania, 29–31 May 2013. doi: 10.1109/CSCS.2013.72

O'Connor, B., Balasubramanyan, R., Routledge, B.R., and Smith, N.A. (2010) 'From tweets to polls: Linking text sentiment to public opinion time series', *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, 23–26 May 2010.

Oliveira, D.J.S., de Souza Bermejo, P.E., and dos Santos, P.A. (2015) 'Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls', *Journal of Information Technology & Politics*. doi: 10.1080/19331681.2016.1214094

Parmelee, J.H., and Bichard, S.L. (2011) *Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public*. Lanham, MD: Lexington Books.

Robert, C., and Rosenbaum, M. (2011) 'A new approach for the dynamics of ultra high frequency data: The model with uncertainty zones', *Journal of Financial Econometrics*, 9: 344–366.

Rothschild, D.M., and Wolfers, J. (2011) 'Forecasting elections: Voter intentions versus expectations'. doi: 10.2139/ssrn.1884644.

Sampson, J., Morstatter, F., Maciejewski, R., and Liu, H. (2015) 'Surpassing the limit: Keyword clustering to improve Twitter sample coverage'. ACM, HyperText conference, Guzelyurt, TRNC, Cyprus, 1–4 September 2015.

Shi, L., Agarwal, N., Agrawal, A., Spoelstra, G., and Spolestra, J. (2012) 'Predicting US primary elections with Twitter', Unpublished manuscript. Available at: http://snap.stanford.edu/social2012/papers/shi.pdf

Silva, N.F.F., Hruschka, E.R., and Hruschka Jr, E.R. (2014) 'Tweet sentiment analysis with classifier ensembles', *Decision Support Systems*, 66: 170–179.

Skaperdas, S., and Grofman, B. (1995) 'Modeling negative campaigning', *American Political Science Review*, 89(1): 49–61.

Topsy Lab (2012) '"The Twitter political index" in Sharp (2012) "A new barometer for the election"'. Available at: https://blog.twitter.com/2012/a-new-barometer-for-the-election

Vargo, C.J., Guo, L., McCombs, M., and Shaw, D.L. (2014) 'Network issue agendas on Twitter during the 2012 U.S. presidential election', *Journal of Communication*, 64(2): 296–316.

Walther, D. (2015) 'Picking the winner(s): Forecasting elections in multiparty systems', *Electoral Studies*, 40: 1–13.

Washington, A.L., Parra, F., Thatcher, J.B., LePrevost, K., and Morar, D. (2013) 'What is the correlation between Twitter, polls and the popular vote in the 2012 presidential election?', Unpublished manuscript.

Zhang, L. (2011) 'Estimating covariation: Epps effect, microstructure noise', *Journal of Econometrics*, 160: 33–47.

4 Leaders, promises and negative campaigning

Digging into an electoral campaign through social media

Introduction

The information available on social media can be useful to answer to a variety of questions related to electoral campaigns. So far, in Chapter 3, we discussed how social media can allow us to nowcast and forecast the evolution of preferences and voting intentions – day by day. The same information, however, can also be exploited to analyze (and nowcast) the effectiveness of campaign strategies, thereby allowing spin-doctors to adjust their strategies almost in real time. Furthermore, social media data can also provide new insights on the study of traditional political sciences topics. In particular, they can answer to questions about the campaign effects focusing on the three main elements that affect the voters' choice: first, the role of political leaders and their valence endowment, in terms of popularity and approval among the voters; second, the impact of policy promises, that is positive campaign related to policy issues made by one party; third, the impact of negative campaigning made by one party against other(s).

Focusing on the evaluation of Italian political leaders (from the autumn of 2012 until the spring 2013) and on the electoral campaign for the Italian general elections (that took place at the end of February 2013), we will see how the analysis of social media can be very informative to detect campaign dynamics and to anticipate trends in public opinion.

As usual, we analyze comments made by citizens and voters to nowcast the campaign and to forecast the final results; in addition, we evaluate, as already stressed, the effectiveness of alternative campaign strategies. To do that, we linked the campaign messages published online by the official accounts of political parties with the changes in the online voting intentions expressed by voters. Finally, we investigate the relationship between the sentiment of political leaders and the voting intentions recorded by survey polls, in the months before and after the campaign.

The online voting intentions in the 2013 Italian election

After the 2012 primary election of the centre-left coalition (that we investigated in the previous chapter), social media acquired a growing relevance in the Italian context. Social media have been widely used by all the Italian political parties and by their leaders in order to carry out the political campaign. They wrote messages to organize and advertise the events of the campaign, such as public rally or television debates, but they also disseminated online the content of such events, publishing all the statements made by their leader during a rally or during interviews with the media. Politicians often used social networking sites to talk to each other, for instance to offer or negotiate alliances or to comment and react in real time to the breaking news of the political agenda. They also replied to the questions raised by the voters.[1] Twitter, in particular, played a relevant role in the following 2013 Italian general election, held in February. The role of Twitter has been so relevant that Mario Monti, who was the incumbent prime minister at that time, announced his choice to run in the election from his Twitter account.[2]

Three main coalitions contested the 2013 Italian general elections: the centre-left was led by Pier Luigi Bersani, selected in December 2012 through open primary election, and composed of the Democratic Party (PD), the left-wing group named Left Ecology and Freedom (SEL) and a small centrist party; the centre-right coalition, led by Silvio Berlusconi, included the People of Freedom party (PDL), the Northern League (LN) and other minor right-wing parties; the centrist alliance was composed of Civic Choice (SC), a new party created in December 2012 to sustain Mario Monti, plus the Union of Christian and Centre Democrats (UDC) and a small splinter group of the PDL. Alongside these coalitions, two other new parties entered the political arena for the first time and ran the election alone: the Five Star Movement (M5S), created by the comedian Beppe Grillo,[3] and Civil Revolution (RC), an electoral cartel composed by Antonio Di Pietro's Italy of the Values, the Greens and two far-left communist parties that selected a former anti-mafia prosecutor, Antonio Ingroia, as their leader.

2013 Italian parliamentary election

Period	16 January–25 February 2013
Data Source	Twitter
Data Gathering	API
Keywords	berlusconi2013, pb2013, italiagiusta, benvenutasinistra, iostoconsilvio, forzasilvio, votaberlusconi, primailnord, sceltacivica, agendamonti, votoudc, pd, sel, centrosinistra, csx, scelta civica, udc, fli, pdl, lega nord, ladestra, fratelliditalia, centrodestra, cdx, m5s, rivoluzionecivile, Fare2013, eleitalia, elezioni2013, elezioni, iovotom5s, monti, bersani, berlusconi, grillo, ingroia, vendola, maroni, casini, padania, SBerlusconi2013, #berlusconi, iolovoto, grandesud, iovotopdl, iovotopd, voterò, voto pd, voto pdl, voto scelta civica
No. comments	7,782,490
SASA Method	iSA
Weights	No
Election day	24–25 February
Forecast release	Daily (public data until AGCOM ban on 8 February; last release: 25 February (15.00, CET)
Reference	http://sentimeter.corriere.it/2013/01/24/elezioni2013-la-politica-vista-dalla-rete/; http://www.corriere.it/Speciali/Politica/2013/elezioni/camera/index.shtml; http://daily.wired.it/news/politica/2013/02/25/risultati-elezioni-previsioni-poll-twitter-65257.html; Civic Choice

The different actors competing in the elections were characterized by specific strategies and messages. The centre-right coalition ran on an anti-tax platform, whose main features were the abolition of the housing tax introduced by the Monti cabinet (IMU), a promise that was championed by Berlusconi's PDL, and the idea, supported by the Northern League, to increase Northern regions' budgets by keeping 75% of income taxes paid by their residents instead of transferring this money to the central government. The centre-left coalition ran a campaign based on values and policies, stressing the need for social justice and equal opportunities ("fair Italy" was the main slogan of the campaign) and the willingness to put an end to Berlusconi's political career. The centrist alliance relied on Monti's record as prime minister to inform a pro-Europe, pro-financial rigor policy platform. The M5S main message throughout the campaign was an invitation to voters to send home all political parties, as a reaction to their alleged corruption and incompetence. Finally, RC presented a left-wing platform and relied on Ingroia's reputation to attract voters.

Once again, we analyzed social media conversations for the whole length of the campaign. From the 16th of January until the 25th of February, when the election was held, the percentage of the voting intentions expressed on Twitter toward the main Italian parties was estimated through iSA. We tried to download the whole population of tweets, written in Italian, commenting on the elections. Tweets were retrieved from public accounts through the Twitter application programming interface (API), using a wide number of keyword search queries (see previous box). Overall, we downloaded 7,782,490 tweets, which was roughly 195,000 per day.[4] Following what was done in the 2012 US election, each daily estimate was measured as a 7-day moving average, taking advantage of the fact that the time series of the analysis was rather long and the daily amount of information was sufficiently large.[5]

We measured the support toward the five electoral coalitions: the centre-left "Italia Bene Comune" (PD and allies), the centre-right (PDL and allies), the centrist alliance "Monti for Italy" (Civic Choice and allies), the electoral cartel Civil Revolution and the Five Star Movement. Within these coalitions, the relative strength of 12 parties was assessed. Inside the centre-left we distinguished three groups, the PD, SEL and "others" (including the Democratic Centre, the Italian Socialist Party and the Südtiroler Volkspartei). Inside the centre-right we separately analyzed the votes share of three groups: PDL, LN and "others" (summing up Brothers of Italy, The Right, Popular Italy and other minor lists). Inside the centrist coalition we accounted for the three parties, SC, UDC and FLI (Future and Freedom for Italy). Finally, we also separately recorded the support of all the minor parties that were not included in these five big groups (almost all were related to the liberal-democratic party "Fare per Fermare il Declino") as well as the share of undecided voters.

In this particular analysis we were, therefore, dealing with a large number of relatively small parties. This represented a further empirical challenge, that of estimating the share of small categories of opinions (i.e. the voting intentions expressed for very tiny parties) which is, indeed, a problematic task, for two connected reasons: it increases the number of categories D to be estimated (see

Chapter 2), and consequently one needs to classify a larger *training set* to obtain a sufficient amount of coding for each category.

Still, when parties (large and small) gather in coalitions as it happened in the 2013 Italian general elections, it is much easier to obtain good estimates of the shares for a coalition than of the shares for each individual party. We exploited this by constraining estimation in the following way (see Ceron et al. 2016). Step 1: the shares of the five main coalitions (plus the two residual categories including all the minor parties and the undecided voters) were estimated using the iSA algorithm (see Chapter 2) using as the target only the coalitions and the Off-Topic category. This gave precise estimates of the coalitions' shares. Step 2: each text that falls in a given coalition was recoded according to the intention to vote for an individual party. Only the subset of text coded belonging to the coalition was reclassified using the iSA algorithm again to produce the constrained distribution of voter share within the coalition. This provided estimates of vote shares for each party of the coalition as well. Step 3 was then replicated for all the coalitions, resulting in a two-stage constrained supervised and aggregated sentiment analysis. Accordingly, iSA is particularly well suited when there is the need to reduce the number of categories, for example coalitions rather than parties.

More precisely, assume a political system with C possible coalitions c_i, $i = 1, \dots C$. Each coalition c_i consists of several parties; for example p_{i1} is the first party in coalition i. In Phase 1, the texts are classified according to coalitions (suppose that we have four coalitions, i.e. $C = 4$), and the proportion of shares of each coalition is estimated using iSA algorithm. Suppose that we obtain $c_1 = 45\%$, $c_2 = 25\%$, $c_3 = 20\%$ and $c_4 = 10\%$. Each text in the training set *tagged* for an individual coalition is hand coded again (this second manual tagging occurs during the coding in Phase 1) according to the parties of that coalition plus the additional category "other coalition". Assume that for coalition c_1, we have three parties. After recoding and iSA analysis, we obtain the following estimates: $p_{11} = 75\%$, $p_{12} = 15\%$ and $p_{13} = 10\%$. In Phase 2, we renormalize these estimates of p_{11}, p_{12} and p_{13} to sum to the share of coalition c_1. Therefore, we set $p'_{11} = p_{11} / c_1 * 100\%$, obtaining $p'_{11} = 33.75\%$, $p'_{12} = 6.75\%$, and $p'_{13} = 4.5\%$. Notice that in this fictitious example, the estimates of p_{13} have been constrained, reducing the bias in the estimation of $P(S = s \mid D = p_{13})$ due to the small amount of texts expressing that opinion in the training set. In addition, notice that given the faint signal for party p_{13} in the whole dataset, the direct estimate of p_{13} combined with all the parties (approximately 18–20) would have led to not only bias but also large standard errors.

The campaign

Figure 4.1 displays the evolution of the aggregate voting intentions for the main coalitions over time.[6] From the top to the bottom, we find: the centre-left (black line); the centre-right (dark gray line); the Five Star Movement, or M5S (gray); the centrist alliance (light gray); and Civil Revolution (very light gray). These data were suitable to analyze once again the dynamics of the electoral campaign as we

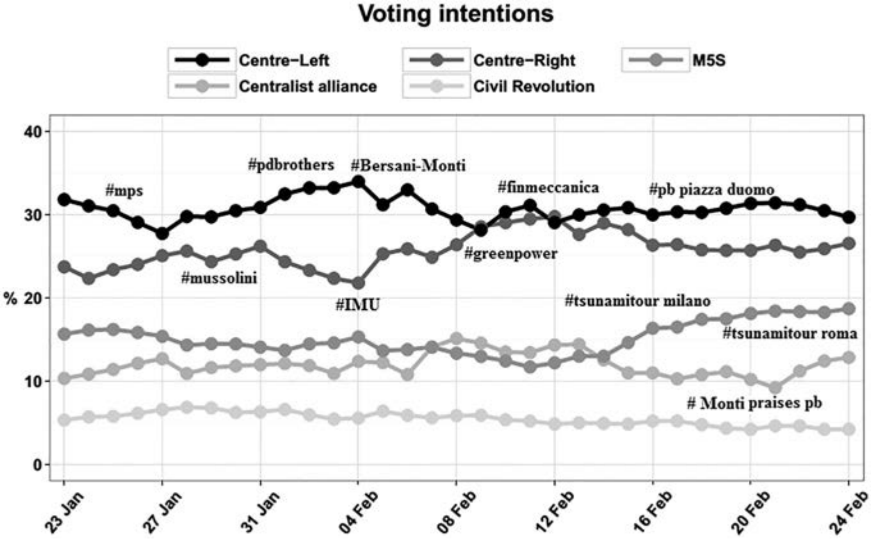

Figure 4.1 Flow of preferences expressed on Twitter (SASA) during the last month of electoral campaign for the five main coalitions in the 2013 Italian general elections

did in the previous chapter. To link the events of the campaign with the changes in voting intention we also reported some hashtags to highlight topics and key events that could have altered the preferences of the electorate.

In detail, we observed that the centre-left was leading almost always, except in mid-February when the growth of the centre-right produced an overtaking. We also observed the striking growth of the Five Stars Movement, in the last 2 weeks before the election, when the *#tsunamitour* intensified. Conversely, in those weeks, the support for the centrist coalition declined. Finally, the left-wing Civil Revolution did not attract much support online and struggled to pass the 4% threshold.

From the graph, we also noticed a decline of the centre-left coalition at the end of January, when a scandal related to the Monte dei Paschi di Siena (*#mps*) bank involved the Democratic Party and seemed to negatively affect its declared voting intention. Indeed, the MPS scandal was so damaging that at the end of January 2013, for the first time, the centre-right coalition approached the front-runner centre-left in terms of voting intentions. However, corruption scandals also hit the centre-right a few days later. In mid-February, the PDL was involved in a scandal related to the Finmeccanica company and this damaged the rating of the centre-right coalition, wielding a decline in the voting intention toward these parties. At first sight, the graph confirms that events of the campaign, like scandals, had, indeed, a negative effect on the expected votes share (Clark 2009; Shaw 1999; Welch and Hibbing 1997). Scandals, however, were not the sole elements that had an impact on the votes share.

The entrance of Renzi into the campaign debate, made in order to respond to the declining vote share of the Bersani's coalition had, indeed, increased the support toward the centre-left (Renzi and Bersani campaigned together, launching the hashtag *#pdbrothers*). We observed a similar (positive) effect also at the end of the campaign, when Romano Prodi, historical leader of the centre-left, entered into the campaign and participated in a rally in Milan to mobilize centre-left voters. Conversely, the first rumors about the creation of a postelectoral coalition between the centre-left and the centrist group led by Monti weakened the consent for the centre-left (*#Bersani-Monti*).

The graph also displays that promises played a role. In particular, the centre-right coalition witnessed a strong increase in its voting intention after its leader, Silvio Berlusconi, promised to give back the house tax (called IMU) to the householder tax-payers, if elected (4 February). In those days, the hashtag *#propostachoc*, used to discuss this proposal, went viral (Cornia 2014). Not all the comments referring to it were positive, and a lot of these were written using ironic language to suggest that such idea was only an unbelievable déjà vu (Cornia 2014). Thanks to our technique, we managed to distinguish criticism and irony from comments of approval. Indeed, we noticed that this proposal was successful in mobilizing centre-right voters, who started to express their voting intentions online at a higher rate, thereby increasing the expected share of votes of the centre-right, closing the gap with the centre-left. This strategy was so effective that – contrary to what emerged from survey polls (Cornia 2014) – in mid-February the centre-right ranked first in terms of online voting intentions. This beneficial effect was observable again at the end of the campaign, when Berlusconi renewed this promise: he sent to voters a facsimile letter (*#letteraIMU*) providing information on how to request to the government to obtain the refund of IMU. This episode may have caused the recovery of the centre-right that was able to almost close the gap in the polls.

As we can see, it seems that campaign events affected the votes share (estimated online) of the different coalitions. This contributes to answering the question of whether electoral campaigns matter for electoral outcomes, which is a heated topic in the literature (Enns and Richman 2013; Finkel 1993; Geer and Lau 2006; Gelman and King 1993; Holbrook 1996; Iyengar and Simon 2000; Lau and Pomper 2002; Shaw 1999; Stevenson and Vavreck 2000; Wlezien and Erikson 2002). Although some authors have argued that elections outcomes are defined in advance (Lewis-Beck and Rice 1992), there has been a more widespread agreement around the idea that campaign events can alter the results (Shaw 1999; Stevenson and Vavreck 2000; Wlezien and Erikson 2002) even though the overall effect could be predictable, minimal or limited to the short run (Gelman and King 1993; Holbrook 1996). This analysis has suggested that different campaign strategies can make the difference. In particular, we observed that corruption scandals and policy promises played a role. However, the effects of campaign events and that of different strategies is more systematically monitored in the next section.

Summing up, the graph sketches some clear patterns. To start with, we observed the rise of the M5S and the fall of the centrist coalition in the last 2 weeks before the election. Such shift was also confirmed by the survey polls. However, what is really interesting here is the fact that the graph describes some trends that were in line with the actual results of the election although they were not revealed by traditional surveys. In particular, we registered the growth of the centre-right coalition whose gap from the centre-left, on average, was smaller in our data compared to the polls. Furthermore, around mid-February, for a couple of days, the centre-right was leading, according to our data. This result seemed reasonable given that the final gap between the two coalitions was below 1%.

The ability of SASA to gauge the narrow gap between the two coalitions and its ability to catch the increase in the expected votes share of PDL after the promise to give back the IMU suggest that, due to the anonymity granted by the Web, online opinions can be less affected by phenomena like social desirability (i.e. the tendency to portray oneself as a "good respondent" or someone whose thoughts, attitudes and behaviors are socially acceptable) and the spiral of silence (Ceron et al. 2014; Noelle-Neumann 1974). In Italy these two aspects are traditionally very relevant, particularly with regard to centre-right voters, who are less willing to express their political view when interviewed for mass surveys and exit polls (Diamanti 2013; Natale 2009). The result is the production of a rather severe measurement error.

Theoretically, people feel more free to express online their personal views without being affected by conformism and social desirability. After all, this is the reason why social media texts have been widely used to study highly sensitive topics such as drug use, sexual behaviors, criminal behavior and controversial political and social issues such as racism and terrorism (e.g. Burnap and Williams 2015; Ceron et al. 2015; Jamal et al. 2015; Zeitzoff et al. 2015). For instance, Berinsky (1999) showed that some individuals who harbor anti-integrationist sentiments are likely to hide their socially unacceptable opinions behind a "don't know" response. Under these circumstances, aggregate public opinion may be a poor reflection of collective public sentiment. Bishop (2004) found similar results with respect to opinion polls conducted on divisive policy issues, such as the teaching of creationism and intelligent design in American public schools. Having said that, we should acknowledge that it is possible to observe the spiral of silence even online, though to a lesser extent (Zerback and Fawzi 2016).[7] If a spiral exists, the supervised sentiment analysis allows us to measure public opinion by taking this aspect into account.[8]

The analysis of the 2013 Italian election is quite interesting also from the point of view of political marketing. Figure 4.2 depicts the estimates of the voting intentions expressed in survey polls toward the centre-left and the centre-right coalitions and reports the gap between the two.[9] For the whole length of the campaign, traditional survey polls suggested that the advantage of the centre-left over the centre-right was quite huge, being around 4–8 points.[10] Furthermore, it was more or less stable (or slightly declining) over time. This overview could lead spin-doctors of the front-runner coalition to adopt a low-profile campaign strategy to avoid risk and preserve the magnitude of the gap. Conversely, the analysis of online conversations highlighted that this was not the case and the race was open. Furthermore, such analysis allowed tracking in real time the effect of public declarations and other events. Accordingly, it represented a new tool (complementary to

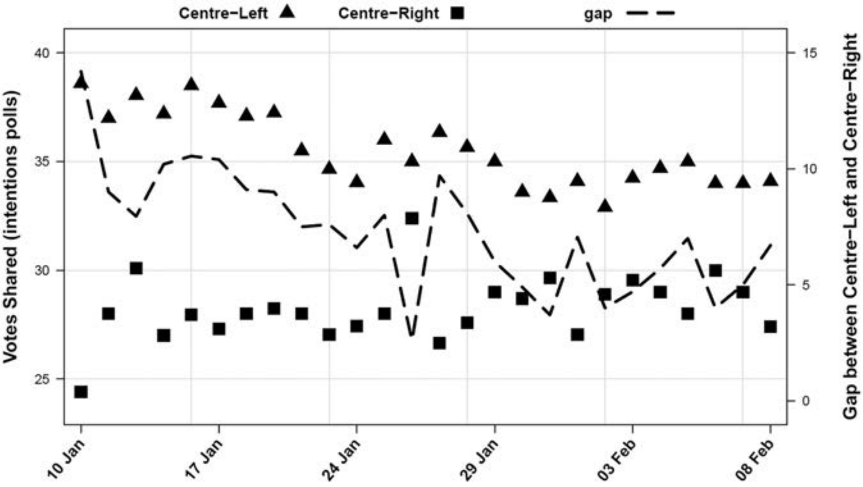

Figure 4.2 Predicted vote share for the centre-right and centre-left coalition according to survey polls throughout the 2013 Italian electoral campaign

traditional surveys) that can be made available to campaign staffs to correct their strategy according to the live reactions of the voters.

Predicting the results

After having nowcasted the whole campaign, our final prediction was released on the 25th of February to be compared with the actual results. The MAE of our estimates was very low: 1.62. Once again, we observed that our aggregate measure of voting intentions was in line with the actual behavior of the electorate and useful to predict the outcome of the election.

In addition, the estimates were strongly correlated with survey polls (*r* = 0.87). When considering all 12 party lists, our MAE was then perfectly in line with that of (clandestine) surveys polls held in the last 7 days before the election, whose average error ranged between 1.26 and 1.86. Surveys, however, made some important mistakes overestimating the Democratic Party and underestimating the M5S. In fact, when considering the aggregated estimates for the main coalitions or macro-groups, our results were markedly better if compared to survey polls. Figure 4.3 displays this comparison. We reported the estimates (and the actual vote share) of four political areas: centre-left, centre-right, centre and the area of “anti”, which sums up three parties that were clearly identifiable as anti-establishment and anti-system: Grillo’s M5S, Ingroia’s RC and Giannino’s (Fare – FID). Indeed, these three parties made appeal to sentiments of anti-politics, protest and renewal. Many tweets posted during the campaign highlighted the similarities between the three parties precisely because they expressed hostility toward the “old politics” represented

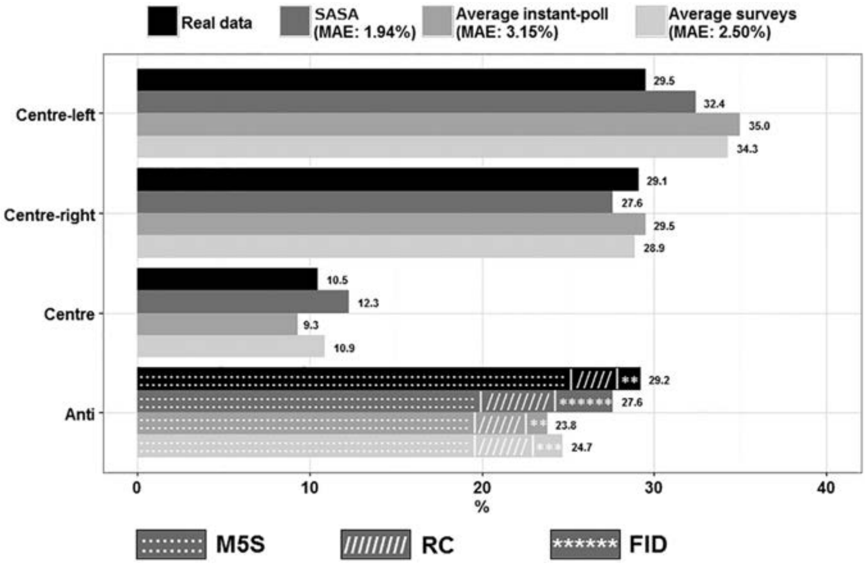

Figure 4.3 Comparison between actual vote share and, respectively, predictions according to survey polls, instant polls and SASA

by mainstream parties. For instance some argued: "*#berlusconi and #bersani are the founding fathers of #M5S! Go #Fare and go #M5S.*" Taking into account these four macro-areas, the MAE of sentiment analysis was still small (1.94%) and it was comparatively much lower compared to that of the last surveys (2.5%) or the instant poll made on Election Day (3.15%).

From this comparison, we noticed that surveys mistook the votes share of the centre-left (overestimated by 5 points) and did not catch the wave of anti-politics, which was underestimated by 5–6 points. This is further evidence that surveys fail to catch the votes share of new parties that run for the first time or that of anti-system parties.[11] In this regard, social media data seemed more suitable to catch such anti-political sentiments (Ceron 2015) expressed by voters who supported the M5S, RC or Fare. This vote of protest was the real surprise of the 2013 election, involving 29.2% of voters. By looking at social media, however, we could have unveiled this surprise in advance. In fact, online "protest voting" intentions amounted to 27.6%, with an error of 1.6 points compared to the actual results.

The effectiveness of positive and negative Twitter-campaigning in the 2013 Italian election

The 2013 Italian general elections and previous Twitter analysis can also be used to measure the campaign effects and to test traditional theories of political science and political communication. In fact, the analysis of electoral campaigns has recently been extended to scrutinize the impact of e-campaigning (see the discussion in Chapter 1). Scholars have initially focused on the politicians' engagement on the Web, showing that it enhances the likelihood of electoral success (e.g. Gibson and McAllister 2011, 2015; Sudulich and Wall 2010). The content of e-campaigning messages though has been rarely taken into account (Wu and Dahmen 2010). The literature usually has classified campaign messages according to their content and has distinguished between negative or positive campaigning. Negative campaigning consists of attacking rival parties/candidates and criticizing their policy platforms or personality traits. Positive campaigning consists of self-promotion messages that emphasize own qualities and proposals.

On the one hand, scholars interested in party competition have posited that parties try to subtract votes to their opponents and, in this light, negative campaigning can be a rewarding strategy. The spatial theory of voting (Downs 1957) suggests that voters' utility decreases along with the distance between their preferred policy program and the platforms proposed by rival parties. Voters will support the party whose policy position is closer to their ideal point. As such, parties with similar electoral platforms will compete against each other. By adopting negative campaigning against its opponent, a party can signal that (all else being equal) the rival party is less suitable to run the country. This will decrease the voters' utility to vote for it, thereby increasing the votes share of the party that enacts negative campaigning.

In multiparty systems, like Italy, negative campaigning is rewarding only when addressed to rival parties with a similar ideological position on the left–right scale (Curini and Martelli 2009, 2015): if a party attacks a rival located far away on the ideological spectrum, voters will not necessarily shift from the target to the attacker because they can choose among several alternatives closer to their own ideological position. Accordingly, they will rather opt for another party, closer to the target on the political spectrum, which is not victim of negative campaigning. For this reason, parties have a strong incentive to attack adjacent rivals whose potential voters are more likely to shift their preferences in support of the attacker (Curini 2015; Walter 2014):

> *H1a: Negative campaigning increases the voting intentions expressed toward the attacker when it is made against a rival party that is adjacent on the left–right dimension.*

On the other hand, some studies have suggested that negative campaigning is an effective strategy because it mobilizes and rallies partisan voters in favor of the attacker (Ansolabehere and Iyengar 1995). In a meta-analytic review of the literature, Lau et al. (2007: 1183) claimed that attacking is not "an effective way to bolster one's own image relative to that of one's opponent". However, they concluded that negative campaigning seems useful to mobilize and stimulate partisans to get out and vote.[12]

The mobilization effect of negative campaign strategies is likely to be particularly relevant in the context of Twitter, due to the hemophiliac nature of social networks (e.g. Conover et al. 2011) and the selective exposure of Twitter users that tend follow and read the content published by like-minded peers (Feller et al. 2011). Therefore, we assume that messages broadcast online will more easily reach "partisan" audiences (i.e. followers) that are more inclined to mobilize after being exposed to negative campaigning.[13] Accordingly, we expect that, by mobilizing partisan voters, negative campaign messages will increase online support (i.e. expressed voting intentions) for the attacker party.

The incentive to adopt negative campaigning, however, need not be constant across the campaign and its impact can be moderated by the general tone of the electoral competition. Going negative could be, in fact, the best reply to an attack (Lau and Pomper 2004). When a party responds to an attack, the backlash effect dwindles since voters will not blame it for defending itself and the party will have an incentive to strike back. As such, the use of negative campaign strategies becomes (increasingly) rewarding when the party is under attack:

> *H1b: Negative campaigning increases the voting intentions expressed toward the attacker when it responds to someone else's attacks.*

Positive campaigning is the other side of the coin. It consists in broadcasting the values, virtues and policies of the party. However, we contend that not all positive messages are equally mobilizing. Taking the cue from the literature on clientelism (Kitschelt and Wilkinson 2007; Piattoni 2001), we split positive campaigning according to its content and focus on messages presenting distributive and clientelistic policies. Promises that involve distributive policies are expected to convey economic benefits to a well-defined subset of voters. As a consequence, when a distributive policy promise is made, its potential recipients have a stronger incentive to react and mobilize online to sustain the fulfillment of such promise by expressing support toward the party making the promise.[14] As such, we argue that distributive and clientelistic policy messages can be, overall, a good strategy to mobilize voters' support, and we analyze their effect net of other positive messages:

> *H2a: Distributive promises increase the voting intentions expressed toward the party making the promise.*

Clientelistic and distributive promises, however, also imply some costs: while targeting promised benefits to certain categories, such electoral claims can alienate the consent of others, like middle class voters (Weitz-Shapiro 2012) or skilled individuals (Calvo and Murillo 2004). Although distributive policies give any party the chance to transfer resources to its constituencies, this strategy helps to mobilize some types of voters more than others; therefore, not all parties benefit equally from them (Calvo and Murillo 2004). Strategies based on distributive promises are more effective when there is a match between the demands of voters) and the ability of parties to meet them through policies (Piattoni 2001):

> *H2b: The effect of distributive promises and its magnitude is conditional on the party making the promise.*

In the 2013 general election, Italian parties made an extensive use of social media, particularly Twitter.[15] They organized and advertised the events of the campaign but also published online all the statements made by party leaders during rallies, interviews or television debates. The analysis of parties' official Twitter accounts then allows us to describe the dynamics of the campaign and to compare the alternative strategies adopted.

To do that, we collected and analyzed all the tweets posted in the official Twitter accounts of the eight largest Italian parties and their leaders: Pier Luigi Bersani (Democratic Party, PD; Twitter accounts: @pbersani, @pdnetwork), Nichi Vendola (Left Ecology and Freedom, SEL; @NichiVendola, @sinistraelib), Angelino Alfano (People of Freedom, PDL; @angealfa, @ilpdl),[16] Roberto Maroni (Northern League, LN; @maroni_leganord, @LegaNordPadania), Mario Monti (Civic Choice, SC; @SenatoreMonti, @scelta_civica), Pier Ferdinando Casini (Union of Christian and Centre Democrats, UDC; @EstremoCentro, @Pierferdinando), Beppe Grillo (Five Star Movement, M5S; @beppe_grillo, @Mov5Stelle), and Antonio Ingroia (Civil Revolution, RC; @AntonioIngroia, @rivcivile).

Per each party and for each day between the 16th of January, when the campaign started, and the 22nd of February, when the campaign ended,[17] we collected and analyzed all 15,053 tweets published on these accounts (396 tweets per day). To perform a time series analysis, these data were aggregated, per each party, on a daily level. Accordingly, the whole dataset consisted of 304 observations (i.e. eight parties monitored for 38 days).

Political messages were hand coded to assess the daily number of meaningful tweets to be included in each of the following categories. To test our hypotheses, we assessed whether tweets performed negative campaign or positive campaign based on distributive promises. Therefore, we distinguished between a) negative campaigning (NC), which records all the attacks made against another leader/party (we also reported the party against which the negative campaign was made) and b) positive campaigning based on distributive/clientelistic promises (PC/Distr) that denotes a message in which the party advocates clientelistic or distributive policies conveying economic benefits to a well-identifiable category or group of voters (e.g. homeowners, unemployed, entrepreneur, taxi drivers).[18] In addition, we also gathered information on two control variables and classified tweets in different categories when they express c) other types of positive campaigning (PC/Others), such as non-distributive policy pledges or any other type of non-clientelistic positive campaign or d) populist statements (POP), that is messages written to criticize the political class as a whole. A number of tweets (around 62%) was related to the organization of the campaign or was written to make generic appeals to support the party without making use of any specific strategy and were not classified in any category.

The work was done by two trained coders. Inter-coder reliability was 0.89. Here are some examples of tweets: "*Berlusconi is always making promises. This is a bad way of doing politics and will damage the country #WithMontiForItaly*" is an example of NC (made by Monti against Berlusconi); the statement "*we will suppress the IMU tax on the primary residence and give back the unfair amount that*

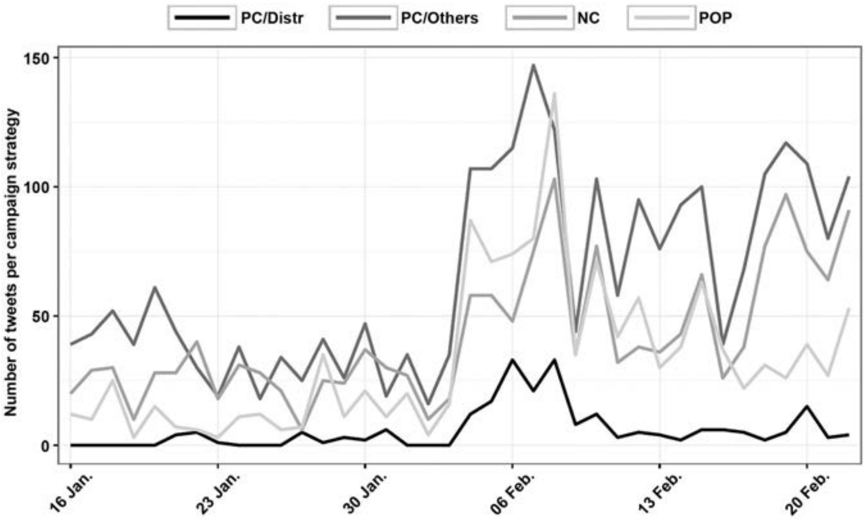

Figure 4.4 Evolution of the different strategies across the campaign

the Italians have already paid" falls in the category PC/Distr, while "*our platform contains my thirty-year experience as entrepreneur*" is considered as PC/Others (both declarations were made by the PDL); finally, "*Those who made such a disaster are now on TV to explain how to get out. They'd better go home #tsunamitour*" is an example of populist tweet, written by the M5S.

The largest share of tweets (16.8%) was written to criticize the opponent, mainly by the PDL, while the party of the incumbent prime minister (SC) reported the lowest level of negativity and this result was consistent with the literature on the propensity to adopt negative campaigning, which was tiny for incumbent parties (e.g. Druckman et al. 2010; Walter 2014). Overall, 19.6% of messages conveyed positive campaigning; here we can distinguish between distributive promises (8.6%), where the PDL, again, had the lion's share, and other positive messages (11%), mostly used by the PD. Only 1.5% of tweets retained a populist content (almost all belong to the M5S).

Figure 4.4 displays the evolution of the different strategies across the campaign. The number of negative campaign messages rose in January, when the PD was involved in a scandal related to the Monte dei Paschi di Siena (MPS) bank and was then criticized by all runner-up parties. Conversely, in the first days of February we observed an increase of tweets advocating distributive policies. This peak seems to have been related to the pledge taken on 4 February by Silvio Berlusconi, who promised to give back the house tax (called IMU) to the home-owning taxpayers, if elected.

We measured the effect of several independent variables related to e-campaigning messages delivered on a certain date (*t*), on the voting intentions expressed via

Twitter in favor of party *i* on the following day ($t+1$). We lost some of the 304 original observations due to the inclusion of the lagged value of the dependent variable.[19]

The dependent variable is the *Share of voting intentions* in favor of party *i*. Residuals categories (that account for minor lists and undecided voters) were excluded; hence the cumulative votes shares of the eight parties included in the analysis do not sum up to 100. Conversely, it varies between 73.5%, in late January, and 90.5%, in the last days before elections. This avoids the problem of dealing with "compositional data" and simplifies the statistical analysis.[20]

The main independent variables, related to the different campaign messages described earlier in this section, are the following: *Negative campaign toward rivals*, which records the number of messages per day written against the closest rival parties along the left–right scale, and *Negative campaign suffered*, that is the daily number of attacks received from rival parties. These were used to test hypotheses H1a and H1b.[21] *Distributive/clientelistic promises* measure the daily number of appeals made to provide distributive and clientelistic policies that benefit a well-identifiable group of voters and were used to test hypotheses H2a and H2b.

To account for the fact that some parties may send more messages than others and for variations in the overall volume of campaign messages, we included other time-varying controls, such as *Party tweets* (the daily number of tweets sent by party *i*) and *Total tweets* (the total number of tweets sent by all parties on the same date). We also controlled for the lagged value of the *Share of voting intentions* and the number of tweets related to other campaign strategies or events such as *Positive campaign*, which includes all the positive campaign messages written each day and unrelated to clientelistic and distributive policy promises, or *Populism*, which accounts for the daily number of populist messages, and *Btp-Bund Spread*, which records the difference in the interest rate paid by the 10-year Italian Btp and German Bundesbank bonds. These two latter control variables capture two topics, populism and the spread, which featured prominently in the Italian political debate over the year preceding the elections.[22]

Empirical analysis and results

Regression results are presented in Table 4.1. Four models are displayed. All the models were estimated through a pooled linear regression, which included party dummies to account for the fact that we have repeated observations nested by party.[23] Note that using different model specifications did not alter the findings. In Model 1 we tested H1a while in Model 2 we included the interaction term between *Negative campaign toward rivals* and *Negative campaign suffered*, to test H1b. In Model 3 we replicated Model 2, adding the interaction between *Distributive/clientelistic promises* and party dummies. Finally, Model 4 included the control variables related to *positive campaign*, *populism* and *Btp-Bund spread*.

Model 1 showed that negative campaign had a positive effect on a party's share of voting intentions (H1a). This result was in line with the existing literature (Curini 2011, 2015; Curini and Martelli 2010) and confirmed that, in a multiparty setting, attacking the a rival adjacent party increased the support for the attacker.[24]

Table 4.1 Effect of different campaign strategies on a party's share of voting intentions

Dependent variable:	*Share of voting intentions (t)*			
	(1)	*(2)*	*(3)*	*(4)*
Share of voting intentions (t − 1)	0.110*	0.096	0.070	0.045
	(0.059)	(0.058)	(0.059)	(0.059)
Negative campaign toward rivals	0.193**	0.022	−0.070	−0.095
	(0.083)	(0.109)	(0.113)	(0.114)
Distributive/clientelistic promises	0.040	0.044	0.119***	0.184***
	(0.039)	(0.038)	(0.045)	(0.051)
Negative campaign suffered		0.030	0.032	0.027
		(0.050)	(0.051)	(0.051)
Negative campaign toward rivals × Negative campaign suffered		0.010**	0.009**	0.009**
		(0.004)	(0.004)	(0.004)
Positive campaign				−0.108**
				(0.054)
Populism				−0.098
				(0.107)
Btp-Bund Spread				−0.004
				(0.021)
Party tweets	0.003	−0.001	0.001	0.009
	(0.009)	(0.009)	(0.009)	(0.011)
Total tweets	−0.001	−0.002	−0.001	−0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Party dummies	YES	YES	YES	YES
Party dummies × Distributive clientelistic promises	NO	NO	YES	YES
Party dummies × Btp-Bund Spread	NO	NO	NO	YES
Constant	11.548***	12.056***	11.932***	13.220**
	(1.229)	(1.267)	(1.261)	(5.765)
Number of Obs	289	289	289	289
R-squared	0.825	0.831	0.840	0.849

Note: Standard errors in parentheses; significance (two tailed): * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Any additional attack made against rival neighbors shifted voters' preferences toward the author of the attack and increased its share by 0.19%.[25]

What is more, the impact of negative campaigning was conditional on the number of attacks suffered, in line with H1b. Figure 4.5 displays the marginal effect of *Negative campaign toward rivals* conditional on the magnitude of *Negative campaign suffered*.

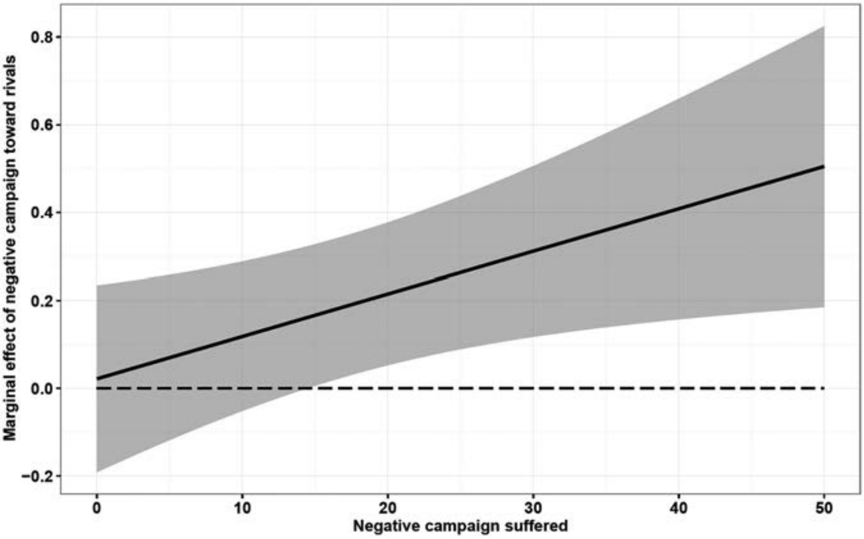

Figure 4.5 Marginal effect of negative campaign toward rivals conditional on the magnitude of negative campaign suffered (with 95% confidence interval)

Negative campaigning increased voting intentions only when the party was considerably under attack and adopted this strategy to answer back. Moreover, the returns from negative campaign were higher the more a party was subject to attacks from other parties. While negativity seemed to matter, we found only a circumstantial effect of distributive and clientelistic policy promises that were effective, though not for all parties, as shown in Figure 4.6.

In fact, the coefficient of the marginal effect tended to be positive for right-of-centre parties while it was negative for the left. However, only the votes share of PD and PDL were significantly affected by distributive promises. The PD was damaged by each announcement (−0.42%) while the PDL had a positive return (+0.12%). Such difference is explained in the next section.

Finally, in Model 4 we showed the robustness of previous results to the inclusion of other control variables (i.e. populism and spread) deemed relevant for the campaign. All the main findings held. Populist messages had no effect,[26] while the spread seemed to play a role though only for two parties: any 10-point increase in the difference Btp-Bund lowered the voting intentions toward the PD by 0.52%, while the M5S share grew by 0.37%.

The statistical analysis revealed that negative campaign seemed to matter (Wu and Dahmen 2010), while positive campaign wielded only circumstantial effects. This was in line with the expectations raised by social psychologists who provided several reasons to explain why negative information could be more persuasive than comparable positive information (e.g. Martin 2004; Pratto and John 1991). On the one hand, people pay greater attention to negative messages rather than to positive

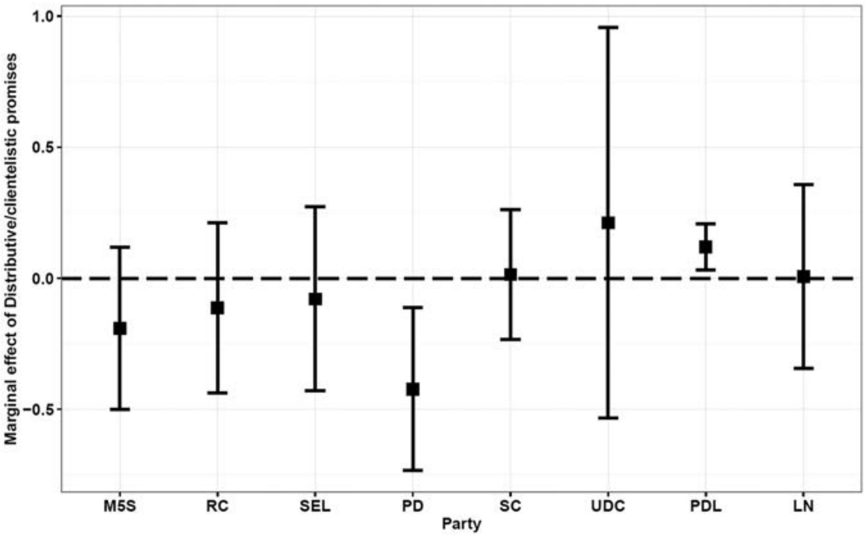

Figure 4.6 Marginal effect of distributive/clientelistic policy promises conditional on the party making the promise (with 95% confidence interval)

ones and the perception of fear generated by negativity can also stimulate interest in the campaign. On the other hand, a negative "flame" also signals to voters that the race is tight and this will bring partisan voters to mobilize and participate.

Our results showed that negative campaign can have a double effect. On the one hand, we confirmed earlier studies (Curini and Martelli 2015) and showed that, in a multiparty setting, negative campaign was effective when targeted against a rival adjacent party, bringing "indifferent" voters (spatially close to both parties) to support the attacker rather than its opponent. On the other hand, we found that the impact of negative campaign was stronger when the attacker was meanwhile under attack: when the party rebutted to an attack, any potential "backlash effect" dwindled because voters would not blame it for defending itself. Being both the source and the object of negative campaign attracted attention and increased a party's prominence in the political agenda, boosting its exposure in the daily debate (online as well offline). Moreover, being under attack forced partisan voters to close ranks and support their party, though we observed this effect only when the party was able to reply, rallying the troops and providing them with arguments useful to answer back. In light of this, we could argue that negativity can also help to mobilize voters who are already ideologically close to the party. If this is true, during the electoral campaign, parties may quarrel against each other and originate flames to activate partisans and sympathetic citizens to get out and vote (Norris 2003; Vaccari 2008). The idea of electoral campaigns as flames could be confirmed by the fact negativity has been beneficial for the two largest poles, PD and PDL, who transformed the campaign in a fight to mobilize support.

Conversely, campaign strategies based on distributive policy promises have only circumstantial effects. In particular, we find differences between the PD, which was hurt by this strategy, and the PDL, which benefitted from it. Why does this happen? A possible explanation deals with the idea that the supply of distributive policies needs to match the demand (Piattoni 2001). The distributive promises related to tax cut plans (including the restitution of the IMU) made by the PDL were perfectly targeted to match its constituency's priorities and mobilize its own voters, while the PD seems to have committed to proposals that were not rewarding for the party. For instance, the Democratic Party promised a reduction of IMU tax as well, but this offer was probably not deemed pivotal by its voters who may have preferred reforms related to welfare spending or tax cuts of another kind. By doing so the PD may also have invaded the area of the political space reserved to the PDL, alienating the support of its partisans (Aldrich 1983), which becomes costly in terms of expressed voting intentions. This result can be explained also by the "issue ownership" theory of voting (e.g. Budge and Farlie 1983), which suggests that voters identify the most credible party that advocates a particular issue and vote for that issue owner, which in this case was the PDL rather than the PD.

It is not surprising that only the PDL benefited from distributive pledges, as other case studies have shown that right-wing parties may have higher returns from patronage (Calvo and Murillo 2004). This is confirmed if we consider that clientelism alienates the consent of middle class and skilled citizens (Calvo and Murillo 2004; Weitz-Shapiro 2012), who are traditionally overrepresented among PD voters (e.g. Maraffi et al. 2013) and underrepresented within the PDL.[27]

Leaders first: the relationship between leaders' sentiment and voting intentions

So far, we have discussed the effect of the electoral campaign and, in particular, the role played by the different campaign strategies on the (online) votes share of each party. However, the literature on electoral behavior has highlighted that nowadays party leaders are one of the most relevant determinants of the voters' choice (e.g. Bellucci et al. 2015). From this perspective, the questions of whether social media analysis can provide informative insights on the relationship between leaders' approval and voters' choice flow out naturally.

We answer the question of the relationship between leaders' appeal and voters' choice, focusing (once again) on the context of the 2013 Italian elections. We analyzed the sentiment expressed online toward political leaders before and after the election. This sentiment was put in relation with time series data on the voting intentions toward political parties recorded by survey polls. From 15 October 2012 until 22 April 2013, we measured at 12 different points in time (every 2 weeks) the sentiment expressed toward 10 Italian political leaders that were the head of the nine main Italian parties (at that time). Given that in the autumn/winter of 2012 the PDL was led by two leaders, that is Alfano and Berlusconi, we accounted for the sentiment of both. Conversely, in 2012 the relevance of the Five Stars Movement (and its leader Grillo) in the political scenario was rather underestimated.

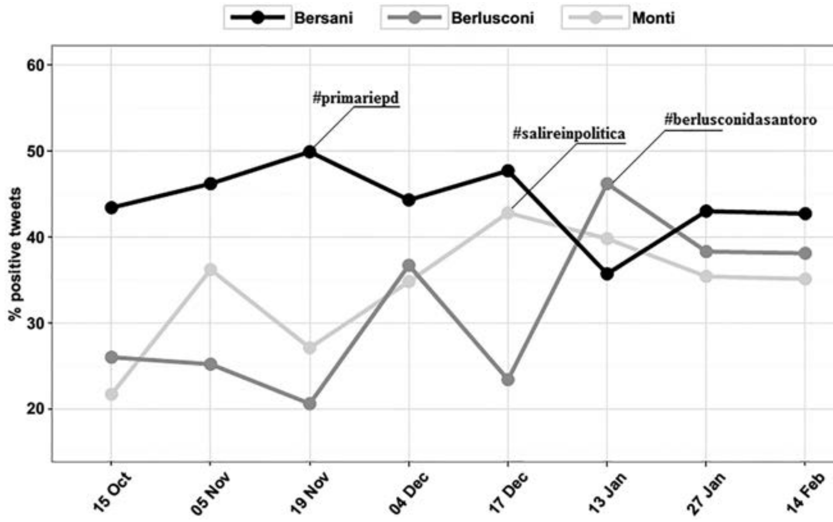

Figure 4.7 Evolution of sentiment toward Italian political leaders before the 2013 Italian general elections (SASA)

Accordingly, we limited the analysis to the following leaders (their respective party is in parentheses): Angelino Alfano (PDL), Silvio Berlusconi (PDL), Pier Luigi Bersani (PD), Pier Ferdinando Casini (UDC), Antonio Di Pietro (Italy of the Values, IDV), Gianfranco Fini (FLI), Beppe Grillo (M5S), Mario Monti (SC), Roberto Maroni (LN) and Nichi Vendola (SEL). Per each leader, by means of SASA, we distinguished whether the sentiment was positive (the comment expressed approval), negative (the comment expressed criticism) or neutral (the comment did not express any judgment).

In Figure 4.7 and 4.8 we report, as an example, the evolution of positive sentiment toward the main Italian leaders before (Figure 4.7) and after (Figure 4.8) the 2013 election. In Figure 4.7 we show the sentiment of Bersani, Berlusconi and Monti, who were unanimously considered as the only leaders who could have potentially become the next prime minister, between 15 October and 14 February. Two weeks before Election Day, the PD leader Bersani ranked first in terms of preferences, with 43% of positive tweets; Berlusconi was second (38.3%) and Monti third (35.4%).

It is worth noticing that Bersani reached the peak of popularity during the (successful) campaign for the 2012 centre-left primary election; in fact, in November 2012, his approval reached 49.9%. To the contrary, in autumn, the approval of Berlusconi was quite low (between 20% and 30%), and it reached the minimum level (20.6%) during the primary election of the centre-left. In those days, in fact, while the centre-left was selecting its leader, the centre-right appeared confused on the theme of the leadership, and there was uncertainty about whether

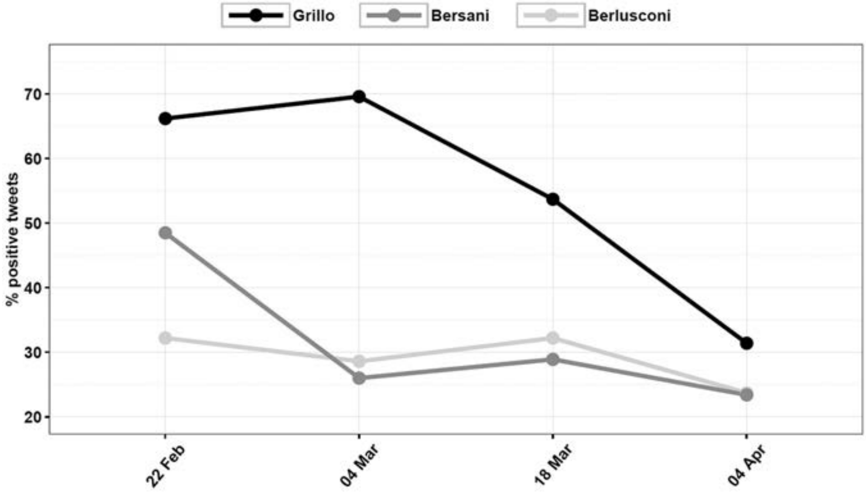

Figure 4.8 Evolution of sentiment toward Italian political leaders after the 2013 Italian general elections (SASA)

the leader would have been Alfano or Berlusconi again. In January, however, the approval of Berlusconi grew as he started to participate to political talk shows on television, particularly during the television talk show *Servizio Pubblico*. This event prompted the approval of Berlusconi up to 46.2%. The link between the appearance of Berlusconi on television and the growth in his approval rate is very interesting and, thanks to sentiment analysis, we can observe that in just one month of campaign he was able to exploit such television shows to boost his popularity by 15 points. The sentiment of Monti is also very informative on the political background. The approval of the former prime minister was low in October (21.7%), and it started to grow only when some rumors anticipated the choice of Monti to run in the election. When he announced this choice, his approval reached a peak of 42.8%. However, when the campaign began, voters noticed that Monti did not feel comfortable as a leader and his approval declined.

In Figure 4.8 we show the sentiment of Bersani, Berlusconi and Grillo (deemed as the most relevant leaders in the aftermath of the election) between 22 February (immediately before the election) and 4 April. In fact, in the aftermath of the 2013 election, these were the three leaders playing a pivotal role in the political scenario, as all the negotiations over government formation included their three parties: PD, PDL and M5S (Ceron and Curini 2014). This graph provides a lot of information on how the sentiment suddenly changed due to the results of the election. Bersani was always the front-runner candidate. However, his failure to win the election, to get a majority of seats and to form a government, dramatically damaged his approval rate. In a few days, he lost around 20 points, and his approval dropped from 48% to around 25%. Conversely, the approval of Berlusconi remained relatively stable. Remarkably, the sentiment of Grillo dropped too. Before the election, Grillo was not perceived as a traditional politician. and he gathered a lot of support online (66.2%). After the striking rise of M5S in the polls, the approval of Grillo grew again after the election, reaching a peak of 69.6%. However, such "honeymoon" period did not last long. From March to April, Grillo's share of positive sentiment was reduced by more than 40 points. Such huge drop can be explained by several reasons. Immediately after the election, the leader of the M5S made some mistakes: he refused to be involved in any negotiation over government formation, and by doing that he proved to be extremely inflexible and not much interested in helping to find a political stability. Other reasons can explain the drop. In particular, after the election results, Grillo started to be perceived as a politician like others. He lost his prevalence on the valence issue (which derived from being an outsider in the political ring) and became subjected to the criticism of opponents, like any other partisan leader.

From this perspective, the mood of social media proved to be relatively unstable as it promptly reacted to different events of everyday politics. Nevertheless, this type of data can be very useful to understand the behavior and the political attitudes of the whole public opinion, as we can observe when contrasting the sentiment toward political leaders with the variation in the voting intentions expressed toward their parties.

In this respect, we assessed whether a variation in the sentiment expressed toward a party leader can affect and anticipate a change in the voting intentions expressed toward his party. To do that, given the large temporal lapse of the analysis (7 months), we gathered data from survey polls, recording the expected votes share of these nine parties. Then we built a simple statistical model to test this relationship by assessing Granger causality. The dependent variable is *Surveys Voting Intentions*, which represents the share of voting intentions attributed, on average, by survey polls to the party *j*, led by leader *i* (in the week *t*). Our independent variable is the lagged share of the *Positive Sentiment* of leader *i* (in the week $t - 1$). We also control for the lagged value of *Surveys Voting Intentions* in survey polls at time $t - 1$. Given that the dependent variable MAE is a proportion whose values are bound by 0 and 1, the assumptions required by the ordinary least-squares (OLS) regression might not hold under this condition due to heteroskedasticity, to the fact that errors might not be normally distributed and the predicted values might fall outside the unit interval (Wooldridge 2002). To address this concern, data were analyzed by means of a fractional logit (Papke and Wooldridge 1996). A fractional logit estimation models directly for the conditional mean of the fractional response through a logistic form that allows us to keep the predicted values in the unit interval (note, however, that using OLS did not alter our findings). Table 4.2 displays the results. In Model 1 we tested this hypothesis, while in model two we double-checked our findings by investigating the reverse relationship: here the dependent variable is the share of *Positive Sentiment* of leader *i* at time *t*, and we included in the model the lagged value of *Positive Sentiment* and the lagged value of *Surveys Voting Intentions*.

Table 4.2 Explaining surveys voting intentions according to leaders' positive sentiment

Dependent variable:	*Surveys Voting Intentions (t)*	*Positive Sentiment (t)*
Positive Sentiment (t − 1)	0.723****	4.177****
	(0.201)	(0.511)
Surveys Voting Intentions (t − 1)	9.041****	0.928
	(0.300)	(0.566)
Constant	−3.487****	−2.547****
	(0.073)	(0.154)
Number of Obs	115	107
Log-Pseudolikelihood	−26.522	−33.030

Note: Robust standard errors in parentheses; Significance (two tailed): * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

The results show that an increase in the amount of positive sentiment at time $t - 1$, increase the voting intentions expressed toward his party at time t. The opposite is not true. When the voting intentions for party j increase, we should not expect a growth in the positive sentiment expressed toward his leader. This finding seemed to suggest that the sentiment toward leaders is a predictor of voting intentions (Bellucci et al. 2015). On the one hand, the strength of the leader has a positive impact on the strength of his party. On the other hand, we noticed that the sentiment seems able to anticipate survey data. Once again, the comments published on social media can become a precious source of information that can be profitably integrated with survey data to enhance our understanding on the evolution of public opinion.

Notes

[1] For example on 5 January, Prime Minister Mario Monti held a question-and-answer session during which he replied live on Twitter to questions posed by his followers (see: http://www.reuters.com/article/2013/01/05/net-us-italy-monti-vote-idUSBRE90408120130105)

[2] See: http://www.corriere.it/politica/12_dicembre_25/monti-napolitano-natale-berlusconi_2dcbfdec-4e9e-11e2-be01–3194f599ff4a.shtml or http://www.lastampa.it/2012/12/26/italia/politica/l-appello-di-monti-su-twitter-basta-lamentarsi-rinnoviamo-la-politica-YP7levPgkFPFBaMIPYXPVK/pagina.html

[3] See Bordignon and Ceccarini (2013) for a description of the history of M5S.

[4] Other sources have agreed that from January until Election Day around 7 million comments were posted on Twitter (http://vincos.it/2013/02/25/elezioni-2013-analisi-di-23-milioni-di-conversazioni-e-interazioni-online/).

[5] The results were made available daily on the online version of the main Italian newspaper, *Corriere della Sera*, until the Italian Authority for Communications Guarantees requested to stop publishing these data in order to comply with the shutdown of opinion polls; in that, the SASA analysis was assimilated to these more traditional techniques of measurement of public opinion. See: http://daily.wired.it/news/politica/2013/02/11/agcom-blocca-voices-from-the-blogs-372825.html.

6 Residuals categories (that account for other minor lists and undecided voters) were excluded, hence the votes shares do not sum up to 100. This simplifies the statistical analysis, avoiding the problem of dealing with "compositional data" (Honaker et al. 2002).

7 See also Burnap et al. (2016) for a work that shows how analyzing social media allows to discount more persuasively the "shy Tory" thesis in the case of the UK elections that has been set forward as an explanation for the mismatch between the prior forecasts and the election outcome: "While there may have been some self-censorship operating on Twitter, the relative anonymity it affords to individuals in expressing their opinions, alongside the entirely voluntary and informal nature of the views expressed we contend would very likely reduce the social desirability bias that affects respondents in a formal survey interview setting."

8 For example if we realize that those who support a particular party tend to express their opinions in a less explicit way, we can classify their comments in a different way to account for this. Indeed, we noticed that the supporters of Berlusconi were less direct in expressing their support for the centre-right. Many comments, in fact, attacked the centre-left while at the same time these texts were written in defense of Berlusconi, though with less explicit statements compared to those used by the supporters of the centre-left. For instance during the MPS scandal, many users claimed that double standards were operating when discussing scandals affecting centre-left (like the MPS scandal) or the centre-right. They complained, arguing that more strictness is used when evaluating the latter. Accordingly, we estimate such tweets as pro-Berlusconi tweets.

9 Data were retrieved from Termometropolitico.it, a website that collects data on polling intentions. We considered the surveys held in the last month before the ban applied to polling companies that forbade publication of data.

10 Notice that, as preelectoral surveys were banned from 9 February, we cannot compare the results of sentiment analysis with survey data following what we did for the 2012 US presidential election.

11 See: http://www.linkiesta.it/blogs/questioni-di-metodo/perche-probabilmente-i-sondaggi-sbaglieranno-parzialmente-le-previsioni-0.

12 Others scholars have emphasized that negativity produces a backlash effect (e.g. Roese and Sande 1993), reducing the affect toward the attacker (but see: Lau et al. 2007; Wu and Dahmen 2010).

13 Twitter users seemed, indeed, more ideologically oriented compared to the whole Italian population (Vaccari et al. 2013).

14 This sends back to the logic of collective action and the idea of social media as a device that facilitates political mobilization by reducing the costs of coordination and participation (Bennett and Segerberg 2011).

15 The incumbent Prime Minister, Mario Monti, announced his decision to run in the election on his Twitter account first.

16 Silvio Berlusconi did not open an official account. We selected Alfano because he was the actual party leader who ran for the premiership.

17 Data cover the whole duration of the campaign. By focusing on the last 40 days before the election, we measured preferences in a period in which shocks were smaller in magnitude compared to those that occurred earlier and the outcome of these shifts tended to persist until Election Day (Enns and Richman 2013; Wlezien and Erikson 2002). Results were robust even when controlling for the number of days before the election.

18 Clientelism, patronage and distributive policies stem from the same process (Calvo and Murillo 2004; Weitz-Shapiro 2012).

19 The inclusion of a lagged term can provide a more stringent test of hypotheses even when, like in our case, not all statistical tests agree that it is necessary to include a lag (e.g. Clark 2013). All results hold when excluding it.

20 Compositional data consist of vectors whose components are the proportion or percentages of some whole and their sum, therefore, is a constant value equal to the whole (e.g. 100%). Votes shares are a typical example.

21 With respect to party *i* we considered as rivals those adjacent parties that do not belong to the same coalition of *i* but are located next to it, on the left or on the right, of the political spectrum. Adjacent rivals are SEL and RC, PD and SC/UDC and PDL and SC/UDC. Since the M5S ran on an anti-system platform, we measured the effect of its negative campaign regardless of the party under attack, as if the M5S were adjacent to all other parties. The results were similar when using placement of parties on the left–right scale suggested in the literature (Ceron and Curini 2014).

22 This electoral campaign was characterized by the presence of populist political actors. In addition, news from financial markets dominated the political debates and was among the causes of key political events, such as the fall of Berlusconi's government and the election of Mario Monti as prime minister.

23 Party dummies control for the idiosyncratic traits of each party such as the leader's charisma, which is assumed constant across the campaign. Coefficients for party dummies and interaction terms related to them were omitted for clarity.

24 The overall effect of negative campaign, regardless of the party it is directed to, was not statistically different from zero.

25 When testing the interaction between *Negative campaign toward rivals* and a set of party dummies, we observed that negative campaigning was more rewarding for the PD (+1.31%) than for the PDL (+0.22%). Conversely, negative campaigning had no effect for the two parties belonging to the coalition that supported the incumbent prime minister, Mario Monti, during the campaign (SC and UDC).

26 Even when testing the interaction with party dummies, populism still had no effect for any of them.

27 Furthermore, the PDL was the only party being overrepresented within the working class (Maraffi et al. 2013), which was more receptive to clientelistic messages.

References

Aldrich, J.H. (1983) 'A downsian spatial model with party activism', *American Political Science Review*, 77: 974–990.

Ansolabehere, S., and Iyengar, S. (1995) *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. New York: Free Press.

Bellucci, P., Garzia, D., and Lewis-Beck, M. (2015) 'Issues and leaders as vote determinants: The case of Italy', *Party Politics*, 21: 272–283.

Bennett, W.L., and Segerberg, A. (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.

Berinsky, A.J. (1999) 'The two faces of public opinion', *American Journal of Political Science*, 43(4): 1209–1230.

Bishop, G.F. (2004) The illusion of public opinion: Fact and artifact in American public opinion polls. Lanham, MD: Rowman & Littlefield Publishers.

Bordignon, F., and Ceccarini, L. (2013) 'Five stars and a cricket: Beppe Grillo shakes Italian politics', *South European Society and Politics*, 18(4): 427–449. doi: 10.1080/13608746.2013.775720

Budge, I., and Farlie, D.J. (1983) *Explaining and Predicting Elections: Issue Effects and Party Strategies in Twenty-Three Democracies*. London: Allen & Unwin.

Burnap, P., Gibson, R., Sloan, L., Southern, R., and Williams, M. (2016) '140 characters to victory?: Using Twitter to predict the UK 2015 general election', *Electoral Studies*, 41: 230–233.

Burnap, P., and Williams, M.L. (2015) 'Cyber hate speech on Twitter: An application of machine classification and statistical modelling for policy and decision making', *Policy & Internet*, 7(2): 223–242.

Calvo, E., and Murillo, M.V. (2004) ‘Who delivers? Partisan clients in the Argentine electoral market’, *American Journal of Political Science*, 48: 742–757.

Ceron, A. (2015) ‘Internet, news and political trust: The difference between social media and online media outlets’, *Journal of Computer-Mediated Communication*, 20(5): 487–503.

Ceron, A., and Curini, L. (2014) ‘The Letta cabinet(s): Government formation and (in) stability in times of crisis—a spatial approach’. *Italian Politics*, 29(1): 143–159.

Ceron, A., Curini, L., and Iacus, S.M. (2016) ‘Using social media to forecast electoral results: A review of state of the art’, *Italian Journal of Applied Statistics*, 25(3): 239–261.

Ceron, A., Curini, L., Iacus, S.M., and Porro, G. (2014) ‘Every tweet counts: How content analysis of social networks can improve our knowledge of citizens policy preferences. An application to Italy and France’, *New Media & Society*, 16(2): 340–358.

Ceron, A., Curini, L., Iacus, S.M., and Ruggeri, A. (2015) ‘Here’s a paradox: Shutting down the Islamic State on Twitter might help it recruit’, *The Washington Post*, December 10. Available at: https://www.washingtonpost.com/news/monkeycage/wp/2015/12/10/heres-a-paradox-shutting-down-the-islamic-state-on-twitter-might-help-it-recruit/

Clark, M. (2009) ‘Valence and electoral outcomes in Western Europe, 1976–1998’, *Electoral Studies*, 28(1): 111–122.

Clark, M. (2013) ‘Does public opinion respond to shifts in party valence? A cross-national analysis of Western Europe 1976–2002’, *West European Politics*. Available at: http://dx.doi.org/10.1080/01402382.2013.841067

Conover, M.D., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., and Flammini, A. (2011) ‘Political polarization on Twitter’. *Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 17–21 July 2011.

Cornia, A. (2014) ‘To refund the property tax: A shocking electoral promise, but also credible? The analysis of survey data and the Twitter debate’, *Comunicazione politica*, 14(1): 97–114.

Curini, L. (2011) ‘Negative campaigning in no-cabinet alternation systems’, *Japanese Journal of Political Science*, 12(3): 399–420.

Curini, L. (2015) ‘The conditional ideological inducement to campaign on character valence issues in multiparty systems the case of corruption’, *Comparative Political Studies*, 48(2): 168–192.

Curini, L., and Martelli, P. (2009) *I partiti nella Prima Repubblica. Governi e maggioranze dalla Costituente a Tangentopoli*. Rome: Carocci.

Curini, L., and Martelli, P. (2010) ‘Ideological proximity and valence competition: Negative campaigning through allegation of corruption in the Italian legislative arena from 1946 to 1994’, *Electoral Studies*, 16: 299–321.

Curini, L., and Martelli, P. (2015) ‘A case of valence competition in elections: Parties’ emphasis on corruption in electoral manifestos’, *Party Politics*, 21(5): 686–698.

Diamanti, I. (2013) ‘Introduzione. 2013: il Paese delle minoranze in-comunicanti’, In Diamanti, I., Bordignon, F., and Ceccarini, L. (eds.), *Un salto nel voto. Ritratto politico dell’Italia di oggi*, ix–xxvii. Rome and Bari: Laterza.

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

Druckman, J.N., Kifer, M.J., and Parkin, M. (2010) ‘Timeless strategy meets new medium: Going negative on congressional campaign web sites, 2002–2006’, *Political Communication*, 27(1): 88–103.

Enns, P.K., and Richman, B. (2013) ‘Presidential campaigns and the fundamentals reconsidered’, *Journal of Politics*, 75: 803–820.

Feller, A., Kuhnert, M., Sprenger, T.O., and Welpe, I. (2011) ‘Divided they tweet: The network structure of political microbloggers and discussion topics’, *Proceedings of the*

Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July2011.

Finkel, S.E. (1993) 'Reexamining the "minimal effects" model in recent presidential campaigns', *Journal of Politics*, 55: 1–21.

Geer, J., and Lau, R.R. (2006) 'Filling in the blanks: A new method for estimating campaign effects', *British Journal of Political Science*, 36(2): 269.

Gelman, A., and King, G. (1993) 'Why are American presidential election campaign polls so variable when votes are so predictable?', *British Journal of Political Science*, 23(4): 409–451.

Gibson, R.K., and McAllister, I. (2011) 'Do online election campaigns win votes? The 2007 Australian "YouTube" election', *Political Communication*, 28(2): 227–244.

Gibson, R.K., and McAllister, I. (2015) 'Normalising or equalising party competition? Assessing the impact of the web on election campaigning', *Political Studies*, 63(3): 529–547. doi: 10.1111/1467–9248.12107

Holbrook, T.M. (1996) *Do Campaigns Matter?* Thousand Oaks: Sage.

Honaker, J., Katz, J.N., and King, G. (2002) 'A fast, easy, and efficient estimator for multiparty electoral data', *Political Analysis*, 10(1): 84–100.

Iyengar, S., and Simon, A.F. (2000) 'New perspectives and evidence on political communication and campaign effects', *Annual Review of Psychology*, 51(1): 149–169.

Jamal, A.A., Keohane, R.O., Romney, D., and Tingley, D. (2015) 'Anti-Americanism and anti-interventionism in Arabic Twitter discourses', *Perspectives on Politics*, 13(1): 55–73.

Kitschelt, H., and Wilkinson, S. (2007) *Patrons, Clients, and Policies: Patterns of Democratic Accountability and Political Competition*. New York: Cambridge University Press.

Lau, R.R., and Pomper, G.M. (2002) 'Effectiveness of negative campaigning in US Senate elections', *American Journal of Political Science*, 46(1): 47–66.

Lau, R.R., and Pomper, G.M. (2004) *Negative Campaigning: An Analysis of U.S. Senate Elections*. Lanham, MD: Rowman and Littlefield.

Lau, R.R., Sigelman, L., and Rovner, I.B. (2007) 'The effects of negative political campaigns: A meta-analytic reassessment', *The Journal of Politics*, 69: 1176–1209.

Lewis-Beck, M.S., and Rice, T.W. (1992) *Forecasting Elections*. Thousand Oaks, CA: CQ Press.

Maraffi, M., Pedrazzani, A., and Pinto, L. (2013) 'Le basi sociali del voto', In Bellucci, P., and Segatti, P. (eds.), *Voto amaro: disincanto e crisi economica nelle elezioni del 2013*, 57–70. Bologna: Il Mulino.

Martin, P.S. (2004) 'Inside the black box of negative campaign effects: Three reasons why negative campaigns mobilize', *Political Psychology*, 25: 545–562.

Natale, P. (2009) *Attenti al sondaggio*. Roma-Bari: Laterza.

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

Norris, P. (2003) 'Preaching to the converted? Pluralism, participation and party websites', *Party Politics*, 9(1): 21–45.

Papke, L.E., and Wooldridge, J.M. (1996) 'Econometric methods for fractional response variables with an application to 401(K) plan participation rates', *Journal of Applied Econometrics*, 11(6): 619–632.

Piattoni, S. (2001) 'Clientelism, interests, and democratic representation', In Piattoni, S. (ed.), *Clientelism, Interests, and Democratic Representation: The European Experience in Historical and Comparative Perspective*, 193–211. Cambridge, UK: Cambridge University Press.

Pratto, F., and John, O.P. (1991) 'Automatic vigilance: The attention-grabbing power of negative social information', *Journal of Personality and Social Psychology*, 61: 380–391.

Roese, N.J., and Sande, G.N. (1993) 'Backlash effects in attack politics', *Journal of Applied Social Psychology*, 23: 632–653.

Shaw, D.R. (1999) 'A study of presidential campaign event effects from 1952 to 1992', *The Journal of Politics*, 61(2): 387–422.

Stevenson, R.T., and Vavreck, L. (2000) 'Does campaign length matter? Testing for cross-national effects', *British Journal of Political Science*, 30(2): 217–235.

Sudulich, M.L., and Wall, M. (2010) 'Every little helps: Cyber campaigning in the 2007 Irish general election', *Journal of Information Technology and Politics*, 7(4): 340–355.

Vaccari, C. (2008) 'From the air to the ground: The internet in the 2004 US presidential campaign', *New Media & Society*, 10(4): 647–665.

Vaccari, C., Valeriani, A., Barberám, P., Bonneaum, R., Jost, J.T., Nagler, J., and Tucker, J. (2013) 'Social media and political communication: A survey of Twitter users during the 2013 Italian general election', *Italian Political Science Review*, 43(3): 381–409.

Walter, A.S. (2014) 'Choosing the enemy: Attack behaviour in a multiparty system', *Party Politics*, 20(3): 311–323.

Weitz-Shapiro, R. (2012) 'What wins votes: Why some politicians opt out of clientelism', *American Journal of Political Science*, 56: 568–583.

Welch, S., and Hibbing, J.R. (1997) 'The effects of charges of corruption on voting behavior in congressional elections, 1982–1990', *The Journal of Politics*, 59(1): 226–239.

Wlezien, C., and Erikson, R.S. (2002) 'The timeline of presidential election campaigns', *The Journal of Politics*, 64: 969–993.

Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wu, H.D., and Dahmen, N.S. (2010) 'Web sponsorship and campaign effects: Assessing the difference between positive and negative web sites', *Journal of Political Marketing*, 9: 314–329.

Zeitzoff, T., Kelly, J., and Lotan, G. (2015) 'Using social media to measure foreign policy dynamics: An empirical analysis of the Iranian–Israeli confrontation (2012–13)', *Journal of Peace Research*, 52(3): 368–383. doi: 0022343314558700

Zerback, T., and Fawzi, N. (2016) 'Can online exemplars trigger a spiral of silence? Examining the effects of exemplar opinions on perceptions of public opinion and speaking out', *New Media & Society*. Advance online publication. doi: 1461444815625942

5 Social media and electoral forecasts

Sources of bias and meta-analysis

Introduction

In Chapters 3 and 4 we have shown how social media can be profitable when used to derive insights on electoral campaigns, to nowcast the campaign dynamics and to forecast the final results. While most of the attempts presented in those chapters were successful, we also observed some variation in the accuracy of the predictions. This is true not only for our analyses (consider for instance the selection of the PD party leader: see Chapter 3) but also for the analyses made by other scholars who have employed alternative techniques (see Chapter 1).

In view of this, the present chapter intends to investigate such variation in the accuracy of the predictions, testing the impact of a number of factors. We start by discussing how the accuracy of the predictions varies over time as information becomes larger and more reliable. For this purpose, two case studies are presented: one is related to the popularity of Italian political leaders in 2011 and the other to the 2015 regional election in Veneto. Later, we exploit data from the 2012 French legislative elections to compare the voting preferences in several local constituencies in order to evaluate potential sources of bias of social media forecasts. Finally, a general meta-analysis of the literature, made by collecting basically all the publically available studies, is performed to explain under which contexts the accuracy of social media predictions increases or decreases, by focusing in particular on the technique used to produce the forecast, the type of election and the amount of information available.

Comparing the approval of Italian leaders in 2011

The first political context in which we try to delve into the factors affecting the accuracy of an SNS 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 Pubblica Opinione) and from the analysis of social media posts. We treated in this sense the former surveys as our benchmark, and we controlled how much the latter approached them.

2011 Italian party leaders

Period	13 January–20 October 2011
Data Source	Twitter
Data Gathering	API
Keywords	Berlusconi, Bersani, Bossi, Casini, Di Pietro, Fini, Vendola
No. comments	31,805
SASA Method	ReadMe
Weights	No
Election day	NA
Forecast release	NA
Reference	Ceron et al. 2014

The mass surveys popularity ratings went from 13 January 2011 to 20 October 2011. They were based on a sample of around 800 respondents, and they focused 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 cofounder 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 ranged from 0 to 100 and identified the percentage of positive scores given by the respondents to each leader.[1]

Similarly, the popularity of each leader from social media was estimated as the percentage of his positive posts over the sum of his positive and negative posts, and once again ranged 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 reran the aforementioned procedure, considering just the week preceding the day in which the mass survey was administered. Overall, we analyzed over 107,000 tweets if we consider the monthly timing (and almost 32,000 tweets in the weekly timing). Given that the results of our analysis looked remarkably similar regardless of the time period considered when analyzing social media, we focus here on the popularity scores that arose from a weekly timing (following the choice made by Tjong Kim Sang and Bos 2012).

In Table 5.1, we report the average difference (mass surveys minus social media popularity ratings) of the scores so obtained. Three main findings clearly arose from the table. First, as long as we consider all leaders without any internal distinction, the average mass surveys ratings appeared to be always higher than social

Table 5.1 Average difference and correlation of leaders' popularity ratings between mass surveys and social media

	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

Note: *n* is the number of mass surveys for each leader

media ratings (on average by more than 5 points). This is true for all the leaders, except Di Pietro and Fini, whose online popularity appeared to be higher. Second, we found a considerable variation among leaders: for example the average difference between the two measures of ratings was quite low for both Bossi and Fini (albeit with a different sign in the two cases), while it increased considerably for Casini, Bersani and Vendola. Third, the correlation between mass surveys and social media ratings was positive, albeit not dramatically strong. Note, however, a marked contrast between, on one side, Berlusconi, Bersani and Bossi (i.e. 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 was, indeed, considerably higher, particularly for Berlusconi ($r = .93$) and Bersani ($r = .75$).

Table 5.1 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 5.1 plots the evolution over time of the MAE of the predictions on leaders' popularity as they arose from social media as compared to the scores obtained from mass surveys. As can be seen, despite being quite relevant at the beginning of 2011 (around 13 points), this absolute difference tended to markedly decrease as times goes by.[2]

To sum up, the first exploratory analysis that we explore in this chapter has provided some quite interesting insights:

1 At least for the most visible leaders, the two measures of popularity ratings (mass surveys and social media) seemed to go hand-by-hand, that is they appeared to react in the same way to exogenous factors (e.g. news reported in the media concerning particular leaders, political events). When the first measure increased, so too did the second one and vice versa.[3]

2 Mass surveys were on average more "generous" than social media with respect to popularity ratings (i.e. they generally gave a higher rating to politi-

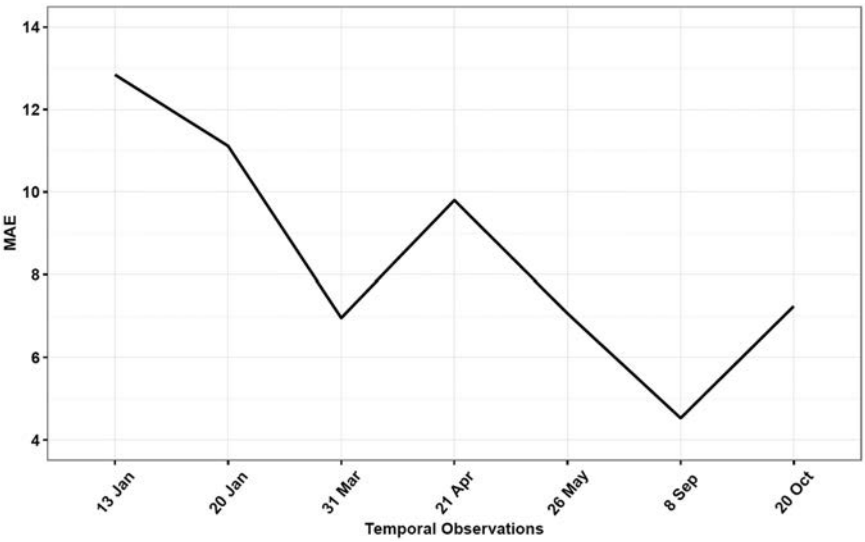

Figure 5.1 Evolution over time of the MAE of social media leaders' approval (SASA)

Note: Mass surveys are used as the benchmark for the prediction.

cal leaders). However, the (absolute) average difference between the two measures of popularity ratings – at least during 2011 – seemed to be clearly declining over time.

In this regard, it is worth noting that the possibility of new elections was widely debated during the second part of 2011 in the Italian political debate throughout the political crisis of the Berlusconi IV cabinet.[4] 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 5.1). This, on the other hand, 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.

Better predictions when the election approaches? 2015 Veneto regional election

The previous example seems to suggest that, as the election approaches, the amount of information available online and the degree of awareness expressed by social media users increases, thereby (possibly) enhancing the accuracy of social media-based electoral forecasts. This argument has been formulated looking at the approval of party leaders in a period that was quite distant from the next election.

Therefore, this claim can be better assessed by looking at predictions made when elections are looming.

2015 Veneto regional election

Period	1 January–31 May 2015
Data Source	Twitter, blogs, forums
Data Gathering	Firehose data provider
Keywords	zaia, alessandramoretti, alessandra moretti, ale_moretti, tosi, m5sveneto, pdveneto, ligaveneta, moretti2015, morettipresidente, facciaafaccia, salvini lega, lega nord, lega veneto, leganord, zaiapresidente, scelgozaia, ilcoraggiodicambiare, iovotomoretti, moretti2015, venetoa5stelle, tosipresidente, flaviotositw, jacberti, berti, veneto2015, veneto
No. comments	481,214
SASA Method	iSA
Weights	No
Election day	31 May
Forecast release	Private reports sent to Veneto Democratic Party
Reference	Veneto Democratic Party

Following this idea, we provide another example in which online opinions have been measured repeatedly over time, both before the beginning of the electoral campaign and during it. For this purpose, we focused on the regional election held in Veneto, in the spring of 2015, called to select the new governor of that region. We started by measuring online attitudes in a period very far from the beginning of the electoral campaign, and we kept measuring them until Election Day.

In that election, there were the two main candidates. One was the incumbent president of Veneto, Luca Zaia (prominent member of the Northern League), and the other was Alessandra Moretti (Democratic Party). We focused on these two and measured the sentiment toward each candidate at four different points in time. The first analysis was done by analyzing all the comments (77,578) written between the 1 January and 28 February, when not all the candidates had officially decided to step in. The second analysis included all the comments (210,597) written between 1 March and 13 April, the third was done between 14 April and 6 May (105,133), finally, the last one was performed using all the comments (87,906) published until Election Day (31 May).

Given that we started the analysis well in advance with respect to Election Day, when voters did not express their voting intentions yet and the names of all the other rival candidates running in the election were not even clear, we decided to focus on the actual sentiment expressed toward the two candidates rather than analyzing voting intentions (as we did in the other forecasts). This different operationalization was done in order to produce results that are consistently comparable over time. Remarkably, while in the early analyses comments tended to express only a general sentiment of approval or disapproval against the candidates, in the last weeks before the election we noticed a growing share of actual voting intentions in favor of one of the two and, for comparative reasons, we considered these voting intentions as positive sentiment.

Accordingly, we recorded the ratio between the positive sentiment expressed toward Zaia and that of Moretti. We noticed that the degree of approval of the incumbent president was always much higher compared to the support of Moretti. In February, when the campaign was not started yet, the advantage of Zaia was quite huge. Conversely, in mid-April our results reported a striking growth in the approval of Moretti. This growth, however, was not sufficient to close the gap between Moretti and Zaia, who was still leading in term of positive sentiment. What is more, the increase in the approval of Moretti can be easily explained by considering that the PD candidate started to campaign early, while the incumbent president was still focused on his administrative duties. As such, the attention devoted to the statements released by Moretti increased and her popularity grew. Such increase was also attested by several survey polls made in those weeks. These survey highlighted that, in April, the gap between the two candidates was on average around 7 points, suggesting that the race could have been close.

Despite such rise in the approval of Moretti, however, the positive sentiment toward Zaia was still higher and, when the campaign started, the gap between the two candidates enlarged once again. In the two sentiment analyses performed in May, in fact, the ratio between the support for Zaia and Moretti confirmed that the incumbent president would easily be reelected.

The MAE of these two predictions was 3.53 and 1.82 respectively, and it was quite low if compared to that of surveys. Furthermore, the error has markedly declined over time as the election date approached. Figure 5.2 reports the mean absolute error of our predictions at the four different points in time.

To the contrary, even in May, survey polls claimed that – although Zaia was leading – the race was still open, and they predicted a good performance of Moretti. In the voting intentions, Moretti ranged between 29% and 35%, according to the surveys. Her expected vote share was overestimated by almost 10 percentage points: the expected value was, on average, 32%, while in the polls Moretti won only 22.7% of votes. To compare the accuracy of survey polls with our estimates, we measured the mean absolute error of surveys (limited to the ratio between the estimates of Zaia and Moretti). Likewise in sentiment analysis, we observed that the MAE of survey polls (slightly) declined over time, passing from 13.6 in April to 10.3 in May. However, the error in this case remained largely higher if compared to that of our social media-based forecasts.

Both the analysis on the approval of Italian leaders in 2011 and that on the attitudes toward the two main candidates running for the governor of Veneto in 2015 have highlighted that, when elections are closer, the reliability of information published on social media is higher. In this regard, such findings confirmed the choice made by many scholars in the field of social media-based electoral forecasts: as we will discuss, around 95% of the social media-based predictions

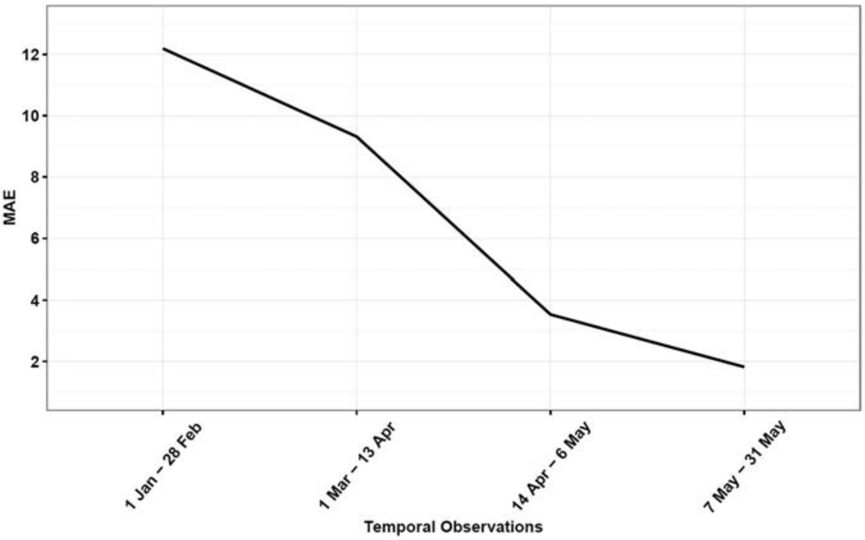

Figure 5.2 The evolution of the MAE over time as the election date approached (SASA)

that we collected in our meta-analysis has been made no earlier than 2 days before the election (see also Gayo-Avello 2013).

The 2012 French legislative elections: when (and where) predictions work

Going beyond these preliminary conjectures, in order to explain variation in the error of the predictions we decided to further scrutinize what elements can affect the ability of social media to forecast electoral results. To do that, we needed to collect a sufficiently large amount of data that allowed us to perform statistical analysis and detect potential sources of bias.

The very first opportunity, in this regard, was provided by another forecast made using SASA to predict the first round of the 2012 French legislative election, held on 10 June 2012. First, we produced a national forecast that was contrasted (as usual) with survey polls and actual votes. Second, we generated a number of predictions using sub-national data, and this allowed us to compare the accuracy of our estimates across 46 local constituencies.

2012 French legislative election (1st round)	
Period	2–9 June 2012
Data Source	Twitter
Data Gathering	API

Keywords	PS, FN, UMP, FDG, EE, EELV, NPA, PG, PCF, Modem, MRC, PRG, CpF, NC, DLR, election, legislatives, legislatives2012, Lutte ouvrière, Nouveau parti anticapitaliste, Front de gauche, Parti communiste français, Parti de gauche, Mouvement républicain et citoyen, Parti radical de gauche, Parti socialiste, Europe Écologie, Les Verts, Mouvement démocrate, Centre pour la France, Front national, Union pour un mouvement populaire, Nouveau Centre, socialiste, gaulliste, verts, NS2012, FH2012, droiteforte, DroiteSarkozyste, gauche, droite, anticapitaliste, extreme gauche, extreme droite, Debout la République
No. comments	79,379
SASA Method	ReadMe
Weights	No
Election day	10 June
Forecast release	9 June
Reference	Ceron et al. (2014)

This case is particularly suitable due to the wide number and the large heterogeneity of local constituencies and to the large number of parties that contested the election; beside seven main parties (such as the Socialists, the Greens, the Ump, the National Front, or the MoDem), we also found a wide number of minor left-wing, right-wing and centrist parties. Overall, we gathered all of them together, based on their ideological affinity, and we measured preferences toward 11 political groups. On the one hand, these features allowed us to compare the ideological orientation of social media conversations covering the whole left–right spectrum, from the extreme left represented by Trotskyist parties, to the extreme right represented by the National Front (FN). As such, it became possible to measure voting intentions with a strong level of detail, testing whether online voting intentions were systematically skewed to the left or to the right or, conversely, moderate large mainstream parties were overrepresented. On the other hand, we could compare our predictions in very different geographical areas, in which left-wing or right-wing parties traditionally gained more votes than their average national share.

At the same time, this forecast represented a harder and more ambitious test for our technique also because supervised sentiment analysis can face troubles when the number of the categories of opinions to be analyzed (in our case, political parties) is too big (as discussed in Chapter 4). To perform our analysis, we gathered 79,300 tweets released during the last week before the elections, and we predicted the national share of votes of the main parties. As shown in Figure 5.3, at the national level, our prediction was once again close to the actual results. This was 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 FN was underestimated. A possible explanation for misestimating the FN vote share is that far-right voters tend to be

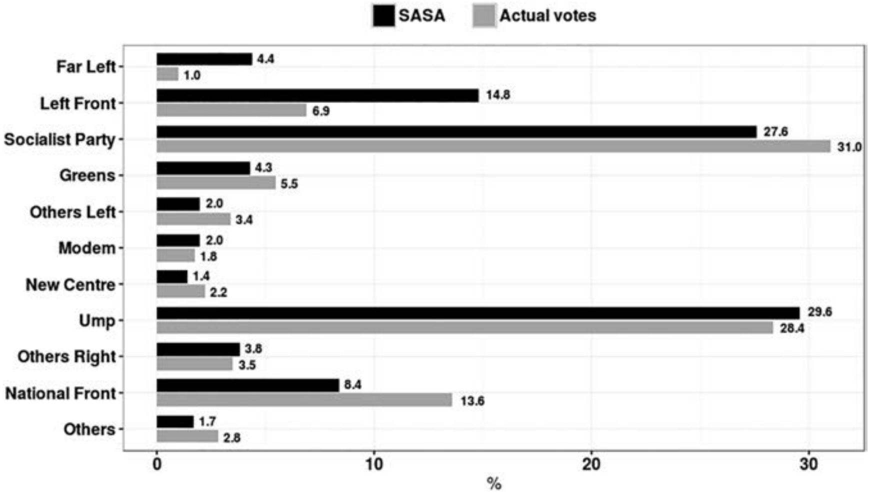

Figure 5.3 Predicted and actual vote shares related to the first round of the 2012 French legislative elections

underrepresented online (this is particularly true for elder voters). In addition, it can be argued that FN voters may be (more) reluctant to publicly express their voting behavior online (notice that even in the first round of 2012 French presidential election, the share of votes of the leader of FN was underestimated). Similarly, left-wing voters seem to be overrepresented in social networks, and this aspect could have led to inaccurate prediction.[5]

That noted, on average the mean absolute error (MAE) of our prediction remained quite low, being equal to 2.38%, which was 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%.

As we highlighted before, these data on the French legislative election allowed us also to explore the main sources of bias that may alter the accuracy of our prediction. To do that, we used 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 analyses to get social media predictions within each area and we compared such estimates with the actual results in the 46 local districts connected to those cities.[6] We measured the MAE, which represented our dependent variable and varied between 2.70 and 8.23. Then we tried to assess which elements increased or decreased the MAE of our prediction. We estimated three different models. In the first one, we included our main independent variables: *Number of Tweets*, the number of comments released in each area, which expresses the information available; and *Turnout*, the percentage of district voters who cast a vote.[7] In Model 2 we added three control variables: *Le Pen Votes Share*, the share of votes gained in the district by the far-right candidate during the 2012 presidential elections (used to identify those areas where the extreme right 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 1 when the incumbent MP is running to seek the reelection. Finally, in Model 3 we added the interaction term between *Number of Tweets* and *Turnout*, to assess whether the effect of having additional information about citizens' preferences was conditional on the likelihood that citizens did actually cast a vote. Data were analyzed by means of fractional logit. Given that we have repeated observations within the same election, we reported robust standard errors clustered by election to avoid possible problems from nonindependent observations or nonconstant variances. Table 5.2 reports the results.

From Model 1 we observed that any growth of the information available online improves our predictive skills. For instance an increase of 1,000 tweets analyzed lowered our error roughly by a quarter point. Conversely, the MAE was greater when the *Turnout* decreased (the same concern affects traditional offline preelectoral polls: see Crespi 1988). This could have happened because some citizens can easily express their opinion online though refusing to cast a vote (perhaps because

Table 5.2 Fractional logit of the MAE

Variables	*(1)*	*(2)*	*(3)*
Number of Tweets	−.000047**	−.0000451**	.00093**
	(.00002)	(.00002)	(.00027)
Turnout	−.0226563**	−.0216124**	.04502**
	(.00961)	(.01059)	(0.125282)
Number of Tweets ×Turnout	–	–	−.000018***
			(4.98e-06)
Le Pen Votes Share	–	.00259	.00093
		(.00701)	(.00678)
Mélenchon Votes Share	–	−.00498	−.00624
		(.01892)	(.01745)
Incumbent	–	−.06816	−.1152
		(.07330)	(.0706)
Constant	−3.677**	−1.404*	−4.969***
	(.4529)	(.7536)	(.9008)
Observations	46	46	46
Log-pseudolikelihood	−7.5809	−7.5796	−7.5675
BIC	−164.5118	−153.0285	−149.2239

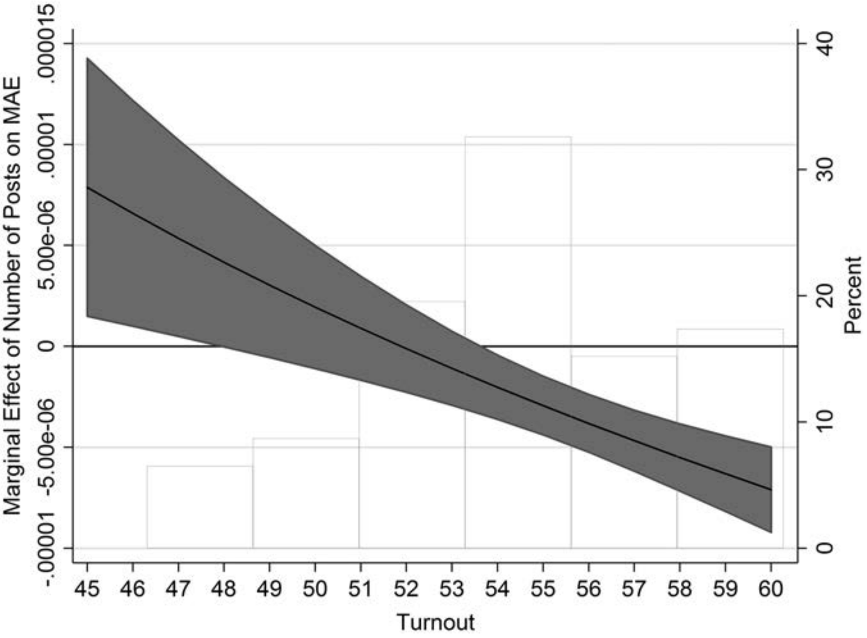

Figure 5.4 Marginal effect of the number of tweets on the MAE as turnout changes (with 90% confidence interval)

Source: Model 3 Table 5.2

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 seemed less able to provide accurate predictions when voters tended to abstain at a higher rate (a 10% decrease in *Turnout* increased MAE on average by 1.2 additional points, while the accuracy seemed greater when turnout is stronger. These two effects held when adding some control variables (Model 2). Our prediction did 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 reelection (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 preelectoral surveys compared to the actual results, due to name recognition. However, Table 5.2 shows that this potential "incumbency effect" did not damage our predictive skills.

Finally, in Model 3, we tested the conditional effect of *Number of Tweets* and *Turnout*. The coefficient of the interaction term is negative and significant. Accordingly, in Figure 5.4 we report the marginal effect of *Number of Tweets* as the level of *Turnout* increases. We also superimpose a histogram portraying the frequency distribution for *Turnout* (the scale for the distribution is given by the vertical axis on the left-hand side of the graph). Despite the relative small substantial impact, it is interesting to note that having more information on citizens' voting preferences decreased in a statistically significant way the error only when the turnout rate was sufficiently high. Up to such threshold, our predictive skills were enhanced by any increase in the number of comments about voting intentions, and such effect enlarged as turnout grew. Conversely, when voters tended to abstain at a higher rate, having more information about their (declared) voting choice negatively affected the accuracy of our predictions: given that voters would express themselves on Twitter instead of casting a real vote, the mean absolute error tended to increase. This (original) finding is quite interesting given that it clearly showed the strict relationship incurring between what happened online and offline in terms of our ability to extract reliable measures from social media analysis.

A meta-analysis of existing studies

To improve the robustness of the findings discussed up to now and to decrease any potential source of selection bias, we tried to maximize the number of predictions considered in the analysis. For this reason, we gathered information on several forecasts based on Web data that have been published in academic journals or that have been presented at academic conferences. We have also considered (until December 2015) those works that are publicly available online and that can be easily found through a Google search using "social media/ Twitter", "election" and "forecast/prediction" as queries.

Scholars have highlighted that works published in academic journals can be subjected to a publication bias: the likelihood of being published seems higher when the paper presents significant results. This implies that scholars would try to submit a paper more often when their prediction is successful (refraining from submitting when it is not), and that these papers might have higher acceptance rates. As such, we observed an inflated number of papers showing positive results (i.e. correct predictions) beyond the actual predictive skills of social media. This bias need not be very problematic thanks to a number of scholars that has supported the alternative argument, trying to provide evidence against the predictive power of social media (e.g. Gayo-Avello 2011, 2012; Jungherr et al. 2012). Nevertheless, to limit this potential selection bias, we decided also to take into account academic works that have not been published as well as nonacademic works that might be less subjected to such concern. The nonacademic predictions represent the 19% of the total.

For the same reasons, when a study displayed several similar predictions we considered only those that were carried out in a substantially different way (different method, different computation of similar methods, markedly different range of dates, etc.). By doing this, we managed to get rid of data whose only purpose was to show the robustness of positive (or negative) results. Analogously, we have excluded "trivial predictions"[8] that might accidentally arise from a bias of the scholars. Following the same logic, we also performed personal computation on available data when the authors did not provide their own measure of accuracy.[9]

Table 5.3 Number of predictions by election type

Type of election	*Number of predictions*
Parliamentary	91
Presidential	80
Local	24
Primary	23
European Elections	15
Referendum	6
Total	**239**

Finally, it has been argued that many forecasts are not actually forecasts because they consist of post hoc analyses (Gayo-Avello 2013). Once again, the fact that data have often been published only after the election may bias the results of our meta-analysis. This further highlighted the need to pay attention to predictions made ex ante, also by considering nonacademic and unpublished works. In our data we reported a reasonably large number of (academic and nonacademic) predictions that were publicly available before the end of the election; these accounted for the 24% of the total number of forecasts.

Overall, we collected and analyzed 239 electoral forecasts related to 94 different elections, held between 2007 and 2015 in 22 countries, covering five continents.

Beyond parliamentary and presidential elections, we also included party primaries held to select leaders and candidates (in France, Italy and the US), local elections (e.g. Alberta, Andalusia, Catalunya, Lombardia and Sicilia as well as Milan, London and so on) as well as the 2014 European elections and the Scottish independence referendum. The large variety of contexts within our dataset (see Table 5.3) was an added value that helped to better investigate the strength and limits of the different methods and to assess which factors can increase (or decrease) their reliability.[10]

On average, we found 2.54 forecasts per election. Almost all the analyses (91%) were done focusing on Twitter alone, and only 5% of forecasts did not consider Twitter as a source of data and exclusively focused on Facebook, blogs, or online news. Analogously, 95% of forecasts were released no earlier than 2 days before the election (as discussed earlier). We noticed a peak in the number of studies related to countries such as Germany, Italy, France, Spain, the UK and the US. The United States was, by far, the country with the highest number of predictions. A wide number of elections (36) held there between 2008 and 2012 has been investigated by several scholars, leading to 79 different predictions. Many of these were related to forecasting the popular vote in the 2012 US presidential election (see Chapter 3). With no fewer than 13 predictions, the race between Obama and Romney (at the national level) has been so far the most studied one. To these figures, we should add seven other analyses of the 2012 US presidential election made at the state level. Beside the presidential race, in the US many other elections were investigated, including many races for the US Senate or for the position of state governor. Primary elections were considered too as well as the election of the House in 2010 and 2012. Italy is the second most studied country. Eleven different competitions were analyzed, leading to 27 social media-based predictions. The most studied were the 2013 general elections, the primary elections of the centre-left and that of PD, but also local elections have attracted the interests of analysts. Analogously, in France and Spain six elections were considered, including primary and local elections. Conversely, in Germany only three different elections (the federal elections in 2009 and 2013 and the European elections in 2014) attracted the interest of scholars, but 22 predictions were made. Finally, beside general elections, in the UK it was the 2014 Scottish referendum that grasped the attention of the scientific community, leading to six different predictions related to it.

Table 5.4 List of countries and number of predictions

Country	*Number of elections*	*Number of predictions*	*Average MAE by country (standard deviation in parentheses)*
Australia	1	4	6.91 (2.65)
Brazil	3	8	7.13 (5.28)
Canada	1	1	9.35 (-)
Ecuador	1	2	7.44 (0.012)
France	6	11	4.12 (3.87)
Germany	3	22	7.11 (4.49)
Greece	2	6	2.68 (1.94)
India	2	5	5.23 (3.41)
Indonesia	1	6	3.22 (3.53)
Ireland	1	8	7.24 (1.38)
Italy	11	27	5.67 (3.85)
Japan	6	7	5.76 (2.77)
Malaysia	1	2	4.03 (2.07)
Netherlands	3	11	2.20 (0.45)
Nigeria	1	4	11.04 (2.97)
Paraguay	1	4	9.83 (7.32)
Portugal	1	2	6.82 (4.64)
Singapore	2	2	5.65 (0.59)
Spain	6	10	2.83 (0.61)
UK	4	14	9.63 (11.76)
USA	36	79	10.30 (7.98)
Venezuela	1	4	6.18 (10.42)
Total	**94**	**239**	**7.39 (6.65)**

For each forecast, we took into account the MAE provided by the authors of the study and whenever such information was missing (though measurable) we computed directly the MAE on available data. Within our sample, the average value of MAE was 7.39, with a considerable variance given that its standard deviation was 6.65. That is some predictions based on social media appeared to be much better (or much worse) than other ones.

Independent variables

Our main independent variable was the method adopted to forecast the election. In line with the literature review, we proposed two baseline models in which we operationalized such variable in two different ways, in order to not only measure the differences between the more general streams of analysis (main approaches) but also distinguish the effect of each specific method (sub-approach) falling within each macro-category (see Table 1.1, Chapter 1).

Accordingly, in the first 2 models we focused on the property of the technique used. In Model 1 we first assessed whether Sentiment Analysis improved the forecasts compared to a merely *computational approach* (our reference category). In this case, the variable *SA* took the value 1 when the analysis was based on automated techniques of sentiment analysis or machine learning; conversely, the variable *SASA* took the value 1 when the forecast was made through supervised aggregated sentiment analysis (*i*SA or ReadMe). To further differentiate within each main approach, in Model 2 we took into account all the sub-approaches displayed in Table 1.1 (Chapter 1). We considered *Endorsement data* (i.e. computation of likes, followers and friends) as the reference category, and we tested which of the remaining five sub-approaches (i.e. the counting of mentions of the party/candidate as captured by the variable *Volume data*; *Traditional SA* that makes use of ontological and predefined dictionaries; *Machine learning*, that is sentiment analysis made through machine learning or building an ad hoc dictionary; *ReadMe* and *iSA*) improved on it.

Given that we were analyzing different elections across political systems, we wanted also to control for a set of potentially confounding factors to exclude any possible bias. First, the variable *Internet users* accounted for the share of Internet users across time and space, measured using World Bank data.[11] This allowed us to control for the pervasiveness of Internet usage in each country. As the share of citizens that has access to the Internet grows, we would expect to observe a lower gap between predictions based on social media and actual results. Under this condition, in fact, the socio-demographic traits of social media users should tend to better approximate those of the whole population of voters, thereby increasing the predictive power of the social media analysis.

In addition, we considered the role of political institutions by distinguishing the effect of different electoral systems. When elections are held in single-member plurality districts (and in majority electoral systems to a lower extent), voters may have an incentive to behave strategically: they can hide their sincere preferences and vote for a candidate who has a significant chance of becoming elected (Cox 1997: see also the discussion in note 5). Conversely, the incentive to behave strategically is lower under proportional representation; hence, the opinions expressed online could be more consistent with the actual behavior at the polls, and this can positively affect the accuracy of social media estimates. We considered this aspect through the variable *Electoral system*, which took the value of 0 for plurality, 1 for majority and 2 for proportional representation.

We also distinguished elections in which voters had to appoint a single person to office (i.e. the president, a party leader, or the mayor) rather than voting for a party list or for a potential member of the national parliament, to assess the possible existence of a "personalization effect" (Mughan 2000) in the electoral campaigning on the accuracy of social media predictions. Accordingly, the variable *Personal vote* took a value of 1 when voters had to select a single politician running for a monocratic position, and it took a value of 0 otherwise.

In Model 3 we included additional variables to test the impact of other particular features of the forecasting techniques adopted. The dummy variable *By user* controlled whether considering only one post per user (rather than considering all the posts published by a single user irrespective of their number) had an impact on the MAE.[12] Counting the number of "users" rather than the whole number of posts refers to the idea that each user has only one vote, so it does not matter how many comments he or she publishes. According to this normalization, a user is considered to favor one candidate if he or she mentions (or likes, etc.) that candidate more often than other candidates (Sang 2012). The variables *Academic* and *Ex ante* controlled respectively whether the MAE was higher or lower for academic papers compared to nonacademic works, and for studies published before or after the end of the election.

Finally, in Model 4, taking the cue from the previous section on the French legislative election case, we included an interaction term between the *Number of posts* (in millions) considered in the analysis (which is a proxy for the volume of information available) and the rate of *Turnout*, to assess whether the effect of having additional information about citizens' preferences was conditional on the likelihood that citizens would actually cast their vote. In fact, we would expect that additional information on the citizens' preferences decreases the error only when the turnout rate is high; conversely, when voters abstain at a higher rate, having more information about their (declared) voting choice would be misleading, as the declared attitude might not coincide with the actual behavior with a negative effect on the accuracy of the prediction.[13]

Because the average error is divided by the number of parties/candidates, this value can artificially lower the MAE and, therefore, we wanted to control for this aspect by measuring the *Number of candidates* considered in the prediction and including this variable in each model.

Given that we have repeated observations within the same election, we reported robust standard errors clustered by election to avoid possible problems from nonindependent observations or nonconstant variance. Table 5.5 displays the results.

Table 5.5 Determinants of the accuracy of social media predictions: fractional logit of the electoral forecast MAE

	(I)	*(II)*	*(III)*	*(IV)*
TECHNIQUE				
SASA	−0.559**		−0.543***	−0.874***
	(0.180)		(0.163)	(0.249)
iSA		−1.444**		
		(0.460)		
ReadMe		−0.774*		
		(0.347)		
SA	−0.168		−0.159	−0.186
	(0.136)		(0.124)	(0.161)
Machine learning		−0.503†		
		(0.299)		
Traditional SA		−0.570		
		(0.359)		
Volume data		−0.402		
		(0.327)		
ELECTION				
Internet users	0.006	0.006(0.004)	0.00502	0.00795
	(0.004)		(0.00507)	(0.00708)
Electoral system	−0.309†	−0.291†	−0.192†	−0.308*
	(0.168)	(0.151)	(0.111)	(0.124)
Personal vote	−0.298†	−0.254	−0.391*	−0.412
	(0.169)	(0.176)	(0.178)	(0.244)
Turnout				0.00208
				(0.00759)
PREDICTION				
By user			−0.151	−0.0822
			(0.133)	(0.160)
Academic			0.0334	−0.206
			(0.227)	(0.261)
Ex ante			−0.240	−0.168
			(0.215)	(0.267)
Number of posts				0.407**
				(0.133)
Number of posts × Turnout				−0.00730**
				(0.00226)
Number of candidates	−0.064	−0.063(0.062)	−0.135***	−0.131***
	(0.069)		(0.0303)	(0.0368)
Constant	−2.255***	−1.896***	−1.924***	−1.931*
	(0.430)	(0.423)	(0.508)	(0.820)
N	236	236	231	165
BIC	−1239.853	−1223.696	−1192.622	−768.0229

Note: Standard errors clustered by election in parentheses

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Discussion

The technique

The results provided strong evidence supporting the idea that the method adopted to analyze the online comments markedly affected the accuracy of social-media-based electoral forecasts. From Model 1 we observed that, all else being equal, the SASA method decreased the MAE by 3.34 points if compared to forecasts based on a mere computational approach and by 2.16 points if compared to other SA techniques, which were not more effective than computational methods in improving the accuracy of the prediction.[14] In Model 2, once again, the two SASA techniques were the only ones that improved on the performance of the reference category (*Endorsement data*), while the impact of *Machine learning* was statistically significant only at the 90% level of confidence. SASA was overall more accurate than any other method. When we further distinguished between the two SASA approaches, we noticed that *iSA* appeared to perform better than *ReadMe*: the predicted MAE of *iSA* (3.01; 95% c.i.: 0.95–5.08) was lower than that of *ReadMe* (5.71; 95% c.i.: 3.78–7.65), therefore, such novel technique further decreased the MAE by 2.70 points.

Election's attributes

With respect to the attributes of the election, we noticed that the *Electoral system* seemed to play a role and its effect was consistent through the different models (although in three models its effect was statistically significant only at the 90% level).[15] Conversely, the *Personal vote* seemed less relevant (a part from Model 3 when *Personal vote* was significant and negative) and the country's share of *Internet users* never affected the MAE. On the contrary, when elections were held under PR, the MAE decreased by a remarkable 2.13 points if compared to plurality. This effect was due to the lower incentive to cast a strategic vote. Because every vote counted in proportional electoral systems, citizens were freer to behave according to their sincere preferences. As a consequence, we observed a higher congruence between opinions expressed online and actual voting behaviors. Conversely, when there was an incentive to behave strategically at the polls, the analysis of the opinions expressed online became less relevant because voters may have expressed their sincere preference online while casting a strategic vote at the polls. This suggested that when some elements prompt the coherence between online opinions and offline behavior, the accuracy of social media-based predictions was heightened. The fact that the error was lower in elections with a high turnout and a huge volume of comments though it was higher when such huge debate was not followed by a large turnout, pointed in the same direction.[16]

Analogously, the interaction between *Number of posts* and *Turnout* pointed to the same conclusion and emphasized the importance of observing a consistency between the opinion expressed online and actual voting behavior. Figure 5.5, based on Model 4, reports the marginal impact on MAE of a one-unit increase in the *Number of posts* – that is an increase by 1 million – as *Turnout* changes

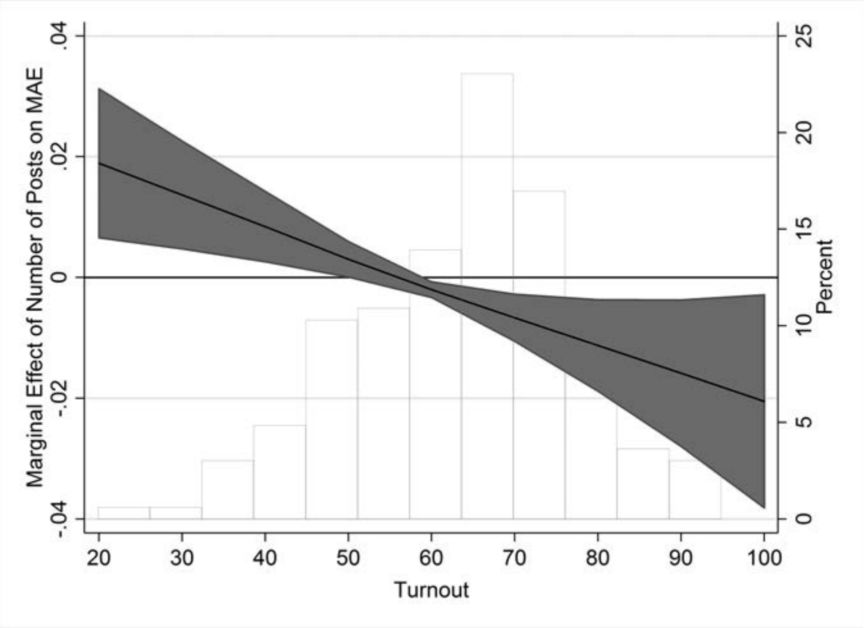

Figure 5.5 Marginal effect of number of posts on the MAE at different levels of turnout (with 95% confidence interval)

its value throughout its range: having more information on citizens' preferences decreased the error, though only when the turnout rate was sufficiently high, that is when we could have expected to observe an actual behavior that was consistent with the declared. The higher the turnout, the higher the positive effects of analyzing a larger number of posts. To the contrary, when turnout was low and voters tended to abstain at a higher rate, they were more likely to express themselves on Twitter rather than to cast a real vote. Here, having more information about their (declared) voting choice negatively affected the accuracy of the forecast, exactly as we found in the previous French case.

Finally, the *Number of candidates* reduced the MAE (as expected). If it is true that, in case of random guessing, predicting the winner of an election is harder when there are more candidates, we must consider that the mean absolute error could be lower because the total error (numerator) is divided by a larger denominator (more candidates).

Prediction's attributes

Surprisingly, whether the prediction came from the academic realm or not, and whether it was made ex ante or ex post did not seem to matter in terms of accuracy, suggesting that – ceteris paribus – these two elements were not significant sources of bias contrary to what we might have expected. The same result applies with *By user*, a result in line with what found in other studies as well (DiGrazia et al. 2013; Tjong Kim Sang and Bos 2012). That is cleaning the data to take into account only one single comment per user, in order to possibly dampen the influence of frequent posters or those with many followers, did not systematically decrease the error. Such adjustments appeared to be not needed or even appropriate. Accordingly concern about "non-representativeness" could also reflect misunderstanding about how prediction from social media analyses works. This is a point on which we will return at length in the next, and final, chapter of this book.

Notes

[1] The survey question on which the popularity rating was based is as follows: "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 or her. If so, please tell me how would you judge him or her, giving a score from 1 to 10: 0 meaning *completely negative judgment*, 10 *completely positive judgments* and 6 *sufficiently positive*."

[2] The peak in MAE registered in April 2011 corresponded to a situation in which the information on leaders' popularity from the mass surveys covered 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 was less than 8 points.

[3] Given the relative scarce time-points available in this analysis, we could not run in this case a lead-lag analysis between surveys and social media results as done in the 2012 US presidential election case discussed in Chapter 3.

[4] 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. 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 clamored for parliament dissolution after Berlusconi's 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.

[5] 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 online, but not at the polls. Albeit in a runoff 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 online and then to vote differently does not exist by definition when we have just two parties/candidates running at the polls. This could also explain why our estimations for the second round of the French presidential election appear slightly better than the French legislative election case.

[6] 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.

[7] Source: French Minister of Interior, http://www.interieur.gouv.fr/

[8] For instance some scholars provided peculiar computations on their data and reported nonsensical predictions in which the votes share of some parties was even negative.

[9] Just to give an idea, some scholars presented their prediction (and the related error) based on a technique of sentiment analysis but from the same source it was also possible to determine the error of other techniques (e.g. computation of mentions).

[10] Note that our substantive findings holds even when we focus strictly on data provided in the literature and when excluding peculiar observations (such as the referendum)

[11] http://data.worldbank.org/indicator/IT.NET.USER.P2

[12] In five cases we were not able to evaluate whether the prediction was made focusing on only one tweet per user as such information was not provided by the source of the forecasts. Accordingly, the number of observations dropped to 231.

[13] Unfortunately, several forecasts did not report the actual number of posts analyzed, therefore, in Model 4 the number of observations dropped to 165.

[14] More in detail, according to the estimates of Model 1 the predicted MAE of a mere computational approach was, all else being equal, 8.25 (95% c.i.: 6.95–9.55); the predicted MAE of other SA techniques is 7.08 (95% c.i.: 5.60–8.55); the predicted MAE of SASA is 4.91 (95% c.i. 3.33–6.50).

[15] Conversely, the type of election (parliamentary, presidential, primary, European, local) did not matter, except for the referendum, which seemed to be less predictable.

[16] Under the idea that when the election is noncompetitive there may be little reason to talk about it (DiGrazia et al. 2013), we also tested the impact of the competitiveness of the election, measured as the margin between the first two parties/candidates or with the effective number of parties/candidates. However, we did not find any effect on the accuracy of the prediction. Data available upon request.

References

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

Ceron, A. (2011) ‘Il governo Berlusconi alla prova delle due fiducie: coesione e divisione tra gruppi parlamentari a fine 2010’, *Polena*, 8(3): 9–33.

Ceron, A., Curini, L., Iacus, S.M., and Porro, G. (2014) ‘Every tweet counts: How content analysis of social networks can improve our knowledge of citizens policy preferences. An application to Italy and France’, *New Media & Society*, 16(2): 340–358.

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.

DiGrazia, J., McKelvey, K., Bollen, J., and Rojas, F. (2013) ‘More tweets, more votes: Social media as a quantitative indicator of political behavior’, *PloS One*, 8(11): e79449. doi: 10.1371/journal.pone.0079449

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

Freeman, S., and Bleifuss, J. (2006) ‘Was the 2004 election stolen’, *Rolling Stone*, 8. Available at: http://www.commondreams.org/views06/0601-34.htm

Gayo-Avello, D. (2011) ‘Don’t turn social media into another “literary digest” poll’, *Communications of the ACM*, 54(10): 121–128.

Gayo-Avello, D. (2012) ‘No, you cannot predict elections with Twitter’, *IEEE Internet Computing*, 16(6): 91–94.

Gayo-Avello, D. (2013) ‘A meta-analysis of state-of-the-art electoral prediction from Twitter data’, *Social Science Computer Review*, 31(6): 649–679.

Gelman, A., and King, G. (1990) ‘Estimating incumbency advantage without bias’, *American Journal of Political Science*, 34: 1142–1164.

Jungherr, A., Jürgens, P., and Schoen, H. (2012) ‘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*, 30(2): 229–234.

Mughan, A. (2000) *Media and the Presidentialization of Parliamentary Elections*. Basingstoke, Hants: Palgrave MacMillan.

Sang, E.T. (2012) Predicting the 2011 ‘Dutch senate election results with Twitter. The Workshop on Semantic Analysis in Social Media’, *Association for Computational Linguistics*: 53–60.

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.

6 Conclusion

"To predict or not to predict?" Future avenues of social media research within and beyond electoral forecasts

Big Data, big mistakes?

The Guardian in 2011[1] revealed that the US government manipulated social media by creating fake accounts with the aim of doing automatic pro-American propaganda. It is not so uncommon to read about VIPs or politicians having fake followers, like Justin Bieber's 37 million followers.[2] By a small error, it is estimated that 83 million Facebook accounts (8.7% of the total accounts) are fake, half of them just duplicates of the same accounts. In March 2014, the *Financial Times*[3] published an article entitled "Big data: are we making a big mistake?", focusing on the big fail of the Google Flu Trends experiment – an example already discussed in Chapter 1 (Butler 2013; Cook et al. 2011) – and many other similar cases. These articles report the great disillusion of the naïve idea that the volume of data itself can replace the scientific reasoning (Anderson 2008), like the enthusiasm (Noble 2003) raised by the early completion of Human Genome Project in 2003, that is the mapping of all genes in the human DNA. Despite having the exact mapping of all genes, those data as-is, have no information per se: once the human genome was transcribed, biologists looked at it with no clue of its real meaning. The same applies to the sea of social network discussions: we can download them all, but they appear to be just noise. Having the data and being able to store, manipulate and move them does not coincide with the ability of extracting information from them.

The findings of the previous chapters have shown what can be considered as the best practice in text analysis, in particular when we are dealing with texts coming from the social media. Such best practices need to satisfy some elementary statistical principles:

1 "Look into the data" or "use supervised statistical methods": data coming from social media contain lot of noise in the sense that, most of the time, the selected texts can be considered as off-topic for the purpose of the analysis. A blind application of ontological dictionaries, the counting of mentions of a candidate (or a party), the count of followers, the number of likes or retweet, are totally biased measures of the willingness to vote for a candidate. Explicit intention to vote must be captured by hand coding of a sample of texts and then this manually tagged set (or the training set as we usually call it) can be used in supervised statistical methods later.

2 Avoid standard data mining techniques in the first place: although designed to discover patterns in the data, they are not designed to extract semantic structure of a text; at most, an average clustering technique may produce aggregation based on the frequency of the words or the like which convey no information for the abovementioned reasons.

3 "Do not perform individual classification": support vector machines, neural networks, random forests, classification trees, logistic or multinomial regressions, and so forth are all based on the assumption that the training set and the test set (the set for which we want to predict the categorization) come from a well-specified population. But this assumption is an unrealistic one. From what was already stated, the highest semantic category in social media data is usually, by far, the off-topic category, but off-topic texts are not as such because they use completely different words from the text of the training set. This means that even the best classifier will attribute with high probability the outcome of the classification to the category off-topic and very rarely the true semantic category. As a result, the individual misclassification will be very high, and the final aggregation of the estimated proportions of the different categories will be highly biased. For this reason, in this particular context, it is better to estimate directly the aggregated distribution of opinions by using new ad hoc method designed explicitly for this field.

When one of the previous pillars of sentiment analysis is missing, the building will risk collapsing into ruins, that is the output of the analysis will be the estimation of the noise (or the bias) rather than the estimation of the signal (Iacus 2014).

Of course, some open issues remain out there. In this conclusion, we focus on two: first, the representativeness of the statistical unit, demographic information of the social media users and so forth. Second, if and how to integrate results from social media analysis with official statistics and standard survey methods.

Do we need to weight the results with individual data?

In the previous chapter, we tested what elements affect the accuracy of the prediction, focusing on the techniques and on the methods used to perform the forecast. So far, however, we have left aside one important element that sends back to a crucial question already mentioned at the beginning of this book.

To forecast an election, we should, in fact, be able to rely on a representative sample, and there is no guarantee that this is something that can be obtained by analyzing social media.[4] On the contrary, socio-economic traits of social media users do not exactly match the actual demographics of the whole population (Bakker and de Vreese 2011; Tjong Kim Sang and Bos 2012; Wei and Hindman 2011) as already mentioned in Chapter 1: people on social media are generally younger (albeit the percentage of elderly people is rapidly increasing)[5] and more highly educated, concentrated in urban areas as well as more politically active overall (Conover et al. 2011; Jensen et al. 2012).

Scholars have tried to apply "debiasing" methods in order to correct any (potential) demographic bias in the Twitter (and in other SNS) user bases according to the voting population (e.g. Choy et al. 2012; Dwi Prasetyo and Hauff 2015; Gayo-Avello 2011; Sang and Johan Bos 2012). These attempts focus mainly on socio-demographic traits that can be automatically inferred from the users' profile, such as gender, age or geographical distribution (Gayo-Avello 2013).

To do that, scholars first gather demographic information about Twitter users. Then the tweets of each demographic group are weighted according to the size of that group in the population of voters (retrieved from National Institutes of Statistics or from data on electoral survey respondents).

In light of this, in Model 1 (see Table 6.1) we have replicated Model 3 of Table 5.5 (as seen in Chapter 5) by adding the variable *Weight*.[6] This variable takes the value of 1 whenever any form of weighting was applied to social media data in order to make the population of users closer to the whole population of voters and to enhance the accuracy of the forecast (Choy et al. 2011, 2012; Shi et al. 2012).[7]

As can be seen, the variable *Weight* was not statistically different from zero (albeit the coefficient has a negative sign as expected). This means that, at least in our sample (and at least so far), weighting the predicted share of votes according to the socio-demographic features of the users, did not appear to improve the MAE of social media predictions (when controlling for all the other explanatory variables). This was a rather surprising result that deserves a deeper discussion.[8]

To start with, we should acknowledge the present limits of the weighting strategies applied to information coming from Big Data. It has been argued that collecting demographic information on the Twitter population is not an easy task and, to some extent, it is not feasible either (Gayo-Avello 2013). While other social network sites (e.g. Facebook) store such information, Twitter profiles do not include structured information and (when such predictions were carried out) Twitter did not ask users to fill in their gender, age or any other socio-demographic trait. Hence, such information could only be inferred indirectly through automated techniques (which can, in turn, generate noise) according to the spontaneous declaration posted by each uses (without any guarantee that, when available, such declarations are sincere or not).

Clearly, the weighting strategies could be improved in the short future. For instance, it has been noted how even extremely nonrepresentative samples can be used as accurate representation of the opinions of a target population with the proper translation of the data into an appropriate indicator (Wang et al. 2014). Still this approach depends on knowledge about the characteristics of respondents included in the sample and a reasonably stable relationship between sample and population. Both preconditions are not met in the case of Twitter (Diaz et al. 2016) as already noted, but perhaps future methodological developments will allow a way.

Having noted that, we must also acknowledge that, to produce an accurate electoral forecast, we should be especially worried about the distribution of political preferences on the Web. Previous analyses have shown, for example that left-leaning is overrepresented, though only marginally (Best and Krueger 2005), while, more recently, it has been shown that Twitter's user base (i.e. the users of the

Table 6.1 Weighting social media predictions: fractional logit of the electoral forecast MAE

	(2)
TECHNIQUE	
SASA	−0.538***
	(0.166)
SA	−0.127
	(0.120)
ELECTION	
Internet users	0.00511
	(0.00487)
Electoral system	−0.194†
	(0.110)
Personal vote	−0.374*
	(0.177)
PREDICTION	
By user	−0.129
	(0.134)
Academic	0.044
	(0.224)
Ex ante	−0.235
	(0.218)
Number of candidates	−0.132***
	(0.030)
WEIGHTING	
Weight	−0.200
	(0.172)
Constant	−1.951***
	(0.496)
N	231
BIC	−1187.231

Note. Standard errors clustered by election in parentheses

†p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

social network most often used for social media electoral forecasts) is ideologically skewed with respect to the population at large (Barberá and Rivero 2014; Huberty 2015; Rainie et al. 2012; Vaccari et al. 2013).

Other studies, however, have found a different result (see, for example Barberá 2015). Table 6.2 compares the distribution of the ideological self-placement of a sample of Italian voters in February 2012 and the subsample that declares to be active on social media. As can be seen, the difference among the two samples is quite trivial.

Table 6.2 Distribution of ideological self-placement of Italian voters versus subsample active on social media

Self-ideological placement	*Whole sample*	*SNS users subsample*
Left	10.50	10.22
Centre-left	15.75	15.25
Centre	14.63	13.81
Centre-right	11.25	11.51
Right	4.13	4.32
None	37.50	38.42
Do not know/did not answer	6.25	6.47

Source: Ipsos survey, February 2013 (Respondents: 800)

Similarly, according to latest Eurobarometer survey that includes variables related to Internet consumption (Eurobarometer 83.2, November 2014), the relationship between self-ideological placement and social network usage over 36 European countries is not significant, neither in a linear way (assuming that leftist people use social network more than rightist people) nor in a quadratic way (assuming that radical people use social network more than moderate ones). The same happens when we explore the relationship between self-ideological placement and those who think that "*online social networks are a good way to have your say on political issues*". In this latter case, the results remain true also when we focus only on those that use social networks quite often (see Figures 6.1 and 6.2, respectively).

Clearly there is a difference between those who use social media or Internet, and those who use social media or Internet and also express a political view (the only ones, after all, who can be captured in an analysis connecting Internet with elections). Having said that, the previous results suggested that we need more analyses before concluding that the distribution of political preferences on the Web are not necessary representative of the political preferences in the population.

It must be also noted that we could have a situation in which those who use Internet and SNS are not representative of the general population, not even from the political point of view, without being necessarily a problem for the accuracy of the electoral forecast. This would happen if those who use SNS tend to go to vote systematically more in a period of decreasing turnout (Franklin 2002). If that happens, they can be representative of the population of actual voters (the things that really matters after all for the forecast), without being representative of the population of the potential voters. The literature on turnout has shown that, among the socio-demographic features, education as well as political knowledge and political engagement increase an individual's propensity to vote (Plutzer 2002; Sigelman et al. 1985). If we look at a recent survey based on the 2013 Italian election (Vaccari et al. 2013), we notice that citizens who discussed the election on Twitter were, indeed, more highly educated, more interested in politics and more

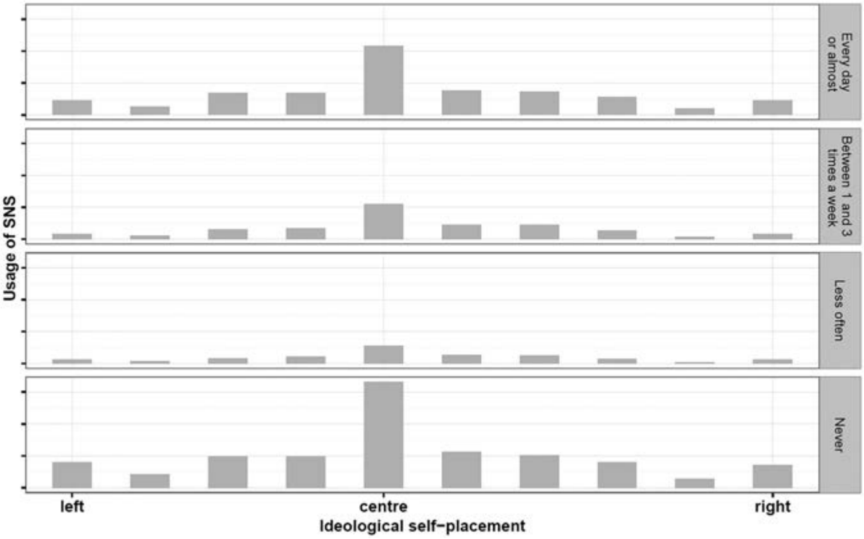

Figure 6.1 Distribution of ideological self-placement of citizens by frequency of SNS usage

Source: Eurobarometer 83.2, November 2014

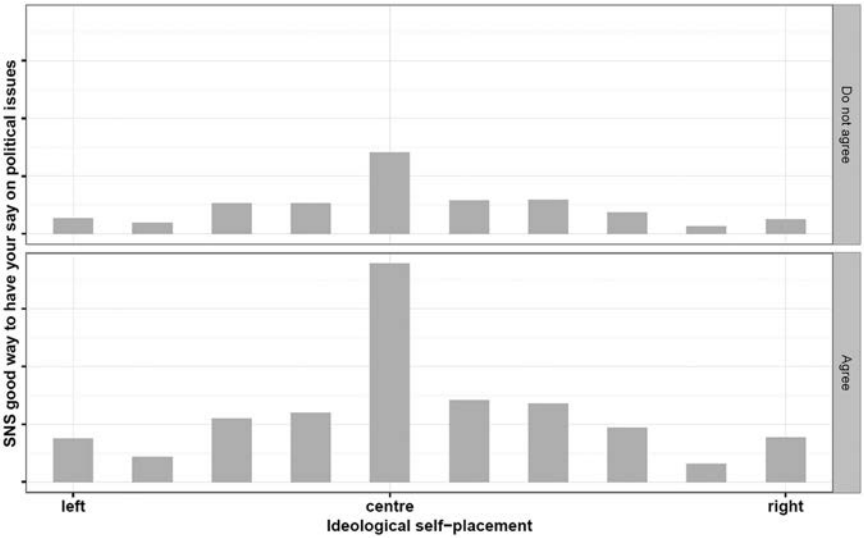

Figure 6.2 Distribution of ideological self-placement of citizens by perceived importance of SNS for expressing own views

Source: Eurobarometer 83.2, November 2014 engaged in offline political conversations if compared to the overall Italian population. These three features suggest that Twitter users can, indeed, be more prone to vote if compared to other people, and this aspect can help explain why Twitter opinions do matter for forecasting purposes. The same holds true even when we generalize beyond Italy. Going back to the already discussed Eurobarometer 83.2, we can find a rather robust correlation between Internet and SNS usage and traditional predictors of turnout such as education, social class, interest in politics and belief in the efficacy of voting (on this point see also de Zúñiga et al. 2012).

Finally, even accepting that the social media population so far is not always representative of one country's population, even of that subset of citizens that casts a vote, there are still some doubts about whether such bias could affect the predictive skills of social media analysis. Indeed, the latter aspect (the predictive skills of social media analysis) does not necessarily require that the previous factor (i.e. the issue of representativeness) holds true. In other words, analyses of social media may capture population-wide distribution of attitudes and behaviors relevant to the topic, even if the characteristics of the user base do not reflect the characteristics of the full population. According to Schober et al. (2016), for example social media data can end up under given conditions to adequately covering the research topics under study, and thus represent the population accurately, even though the individuals who contributed to the social media corpus are not sampled in a representative way.[9]

The ability to achieve topic coverage (whether through population coverage or not) can happen, for example if we assume that Internet users act like opinion makers who are able to influence (therefore often anticipating) the preferences of a wider audience (O'Connor et al. 2010), including the ones of the broader media ecosystem (Farrell and Drezner 2008). The fact that quite often journalists are among the most active consumers of social media (Lasorsa et al. 2012; Spierings and Jacobs 2014) could provide an empirical ground to this hypothetical claim. Similarly, Herbst compared the talk of the Internet to the 18th-century salons: "the conversation of the salons reflected and shaped the culture of France and much of Western Europe and ignited the revolutions that would change our world forever. We do not take the salons lightly now; they are invaluable to historians. And we should treat the internet in precisely the same way" (Herbst 2011: 95). The same applies if social media discussions are able to reproduce all the (more general) public opinions. For example this can be true if Twitter communications are considered to function like a critically engaged interaction system (Ampofo et al. 2011) whose communications about specific issues (such as electoral contests) are thematically representative of larger streams of conversations and preference distributions (Jensen and Anstead 2013).

Having said that, the nonrepresentativeness of social media users is probably more challenging for analyses that focus on the individual opinion of each single user, rather than on the whole amount of comments in the aggregate.[10] On one extreme, as we have shown, we find analyses in which scholars select only one comment/tweet per user. They evaluate his or her preference through that and sum up the preferences of all the individuals to estimate the final outcome (which can, in turn, dramatically decrease the accuracy of the estimation, as we saw in Chapter 2); in this case, they are implicitly assuming that the opinions of those who express an opinion online are correlated with the distribution of opinions offline. If this is not the case, and such mismatch is not weighted, then we can expect to observe a bias in the prediction. On the other extreme, we can find analyses based on aggregated estimates (such as SASA). By working on the aggregate, such techniques are not appropriate for applying weights based on the demographic features of individuals (a possible weakness of such methods). Nevertheless, these methods, precisely because they focus on the aggregate, can be in a better position to recover the "topic coverage" discussed earlier.

On top of that, the nonrepresentativeness of social media users is also less problematic for nowcasting the elections and for analyzing the effectiveness of campaign strategies. In both cases, in fact, we want to pay attention to the (spontaneous) reaction and the process mobilization/de-mobilization of voters rather than in creating a representative sample. Accordingly, our quantity of interest should be the percentage of unsolicited voting intentions expressed online (see, for instance Chapter 4) as a function of the campaign strategies and of the campaign messages broadcast by parties and candidates (when we want to evaluate their effectiveness) or, more in general, the campaign events (when we want to nowcast the campaign). By focusing the attention on these swings we can try to make inference on future trends.

Opinion polls and social media analysis: what relationship?

A second issue that deserves to be discussed involves the relationship between opinion polls and online sentiment. The usage of social media in conjunction with surveys has followed till now two main patterns. The first one treats social media as a way to supplement the survey process through questionnaire development, recruitment and locating (see Moy and Murphy 2016). For example social media platforms have been used for direct recruitment of nonprobability samples for surveys (Bhutta 2012; Rhodes and Marks 2011) as well as to target, recruit and even conduct self-guided online focus groups (see Murphy et al. 2014).

The second approach considers social media as a supplement to traditional survey research methods, or vice versa. As an example of the former possibility, Murphy (2014) surveyed individuals and then supplemented their responses with information gleaned from the content they posted on Twitter. The latter possibility, on the other side, is more in line with the content of the present book. Several recent social media-based predictions (Franch 2013; Tsakalidis et al. 2015), in fact, have tried to encapsulate data from traditional survey polls, building a bridge between these two domains of public opinion studies in order to foster the accuracy of the forecasts. In particular, these techniques try to use social media data to reproduce the time series of survey polls thereby forecasting voting intentions on Election Day. More in detail, Tsakalidis et al. (2015) incorporated Twitter data into poll-based regression models in order to predict election results; the authors gathered information, on a daily basis, on the mentions and the sentiment toward parties and political leaders, generating different time series based on such data. In the meantime, they also collected opinion polls data from different sources to create analogous time series of voting intentions toward each party. Then Twitter data and surveys data were normalized and were put into a regression model, which tries to forecast the values of the polls for the next day. The prediction related to Election Day was the one that corresponds to their forecast.

We have, therefore, replicated the analysis shown in the previous Model 1 (see Table 6.3) by also adding the variable *Polls*, which controlled whether weighting social media data according to the estimates of traditional survey polls had an effect. In addition, we also wanted to assess how difficult it was to predict the election based on information provided by preelectoral survey polls. Accordingly, we have included the control variable *MAE Polls*, which is equal to the mean absolute error of surveys published in the last month before the election.[11] The higher the *MAE Polls* variable, the lower the predictive power of survey polls. As such, we can assess whether social media-based predictions were more (or less) accurate in contexts in which the errors of predictions made focusing on survey polls higher or lower.

Before discussing the results of Table 6.3 more in depth, we must observe that preelectoral opinion polls were – on average – more accurate than sentiment analysis in predicting the final outcome of the elections. Overall, the MAE of opinion polls was, in fact, equal to 2.22 in our sample and, therefore, more than three times lower than social media analysis (an average MAE equals 7.39 as already noted in Chapter 5). Still, although opinion polls still appeared to be on average a better predictor of voting intentions, social media analysis can perform as well as surveys, depending on the technique used to analyze the sentiment of the Web. Furthermore, we can find examples of elections in which social media have outperformed preelectoral survey polls.

In more than 40% of cases, the accuracy of social media predictions was, in fact, in line with survey polls (with a limited difference, lower than 2 points). Moreover, approximately 1 social media-based prediction out of 4 outperformed survey polls. The range of predictions that outperformed survey data was quite heterogeneous. This list included presidential elections held in non-Western countries such as Brazil, Indonesia and Venezuela. But the same happened with respect to the 2014 European elections in Greece and Germany, or with respect to the Scottish referendum and the UK parliamentary elections in both 2010 and 2015. Several predictions related to the popular vote in the 2012 US presidential also outperformed preelectoral opinions polls too. What is more, many of the predictions based on SASA were more accurate than surveys; this happened in the already mentioned 2012 US presidential and in the Scottish referendum, but also in the 2012 centre-left primary, in the regional elections held in Veneto (2015) and Lombardy (2013) as well as in 2010 US midterm elections (Nevada Senate, Alaska Senate, Massachusetts Senate and Massachusetts Gubernatorial elections). This latter case is particularly intriguing, given that the 2010 US elections have been highlighted as an example to suggest that social media cannot be used to predict elections (Gayo-Avello et al. 2011). While it is true that not all the predictions related to the 2010 midterm elections were accurate and successful, it is also

Table 6.3 Survey data as determinants of the accuracy of social media predictions: fractional logit of the electoral forecast MAE

	(2)
TECHNIQUE	
SASA	−0.540*** (0.158)
SA	−0.130 (0.124)
ELECTION	
Internet users	0.001 (0.005)
Electoral system	−0.227* (0.103)
Personal vote	−0.419* (0.166)
PREDICTION	
By user	−0.208 (0.133)
Weight	−0.224 (0.167)
Academic	0.084 (0.212)
Ex ante	−0.036 (0.229)
Number of candidates	−0.121*** (0.029)
SURVEY-RELATED	
Polls	−1.371*** (0.337)
MAE Polls	−0.098* (0.049)
Constant	−1.496* (0.594)
N	215
BIC	−1075.948

Note: Standard errors clustered by election in parentheses

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

true that in many cases (around 1 in 2) such predictions were accurate (in 11 cases the MAE was lower than 5), and five (based on mentions, sentiment analysis or SASA) were even more accurate than survey polls.

Does this mean that social media analysis can be better than surveys polls? Our analysis has suggested that, so far, this is not (always) the case. To the contrary, we can notice from Table 6.3 that anchoring social media predictions to surveys can dramatically decrease the MAE and, therefore, social media data can be profitably integrated with survey data to increase the performance of the prediction. The MAE decreases, in fact, by 5 points on average when surveys (i.e. the variable *Polls*) are taken into account when doing social media electoral forecasts.

Interestingly, we have found also a negative correlation (−0.11) between the MAE obtained by traditional survey polls, which summarized how the election was easily predictable using traditional tools, and the MAE of social media predictions. This result is confirmed in our meta-analysis (see Table 6.3), which revealed that the MAE of social media was lower in electoral contexts in which the MAE of surveys was higher.[12] In other words, the MAE of social media analysis and that of traditional survey polls went in different directions. When the MAE of the survey polls increased, that is when electoral results could be hardly predicted by means of surveys, the MAE of social media analysis, that is the accuracy of social media analysis decreased. And vice versa.

Why did this happen? Probably the information available on social media allowed to catch some trends that were not captured by traditional opinion polls; for instance, voting intentions toward new parties, anti-system parties or parties that were markedly affected by the spiral of silence due to social desirability. These features as well as other trends that manifest themselves in the online realm first (see, for instance Chapter 3), may not emerge when analyzing public opinion through traditional techniques, thereby increasing the MAE of surveys; however, these same trends might be caught by social media analysis, and for this reason, the MAE of social media predictions can decrease.

This result pointed, therefore, to the importance of integrating social media data and survey data. If it is true, as our analysis has shown, that survey data can improve the accuracy of social media predictions, it is also true that survey research can benefit from listening to social media-based predictions, as these predictions in some cases can anticipate new trends and can be used to evaluate changes in public opinion and to adjust the estimates of surveys.

Summing up, to reiterate what was already noted, our review and the connected statistical analysis suggested that the analysis of social media was not always so bad as it was sometimes depicted and, in fact, – depending on the technique used or on the electoral context – we observed many promising predictions based on social media. In this regard, this is a way that is worth pursuing.

Conclusion

The application of sentiment analysis with respect to politics and electoral results is just one of the many examples employing Big Data we have witnessed in the recent years. Some of those examples have clearly shown the utility of addressing old research questions in new ways, and addressing new questions that emerge from this process. After all, the goal of social science research is to understand causal processes and this inevitably requires deductive theory testing (Dalton

2016). Social and political scientists are accustomed to using statistical methods to test theories. While doing it, the focus is typically on the coefficients for the theoretically-derived input variables. When using Big Data in politics (in particular with respect to elections), the focus is usually on the outputs rather than the inputs (Wilkerson and Casas 2017), leading researchers to be first and foremost concerned with prediction accuracy and less concerned with explanation. Still, there is nothing inherently contradictory between Big Data methods and theory testing (Nagler and Tucker 2015). For example, in Chapter 4 we have shown what we can learn by employing social media data in terms of the role of political leaders and their valence endowment, the impact of policy promises and that of negative campaigning. Of course, other researchers highlight the limits of these new types of data.

This is far from being a surprise and, on the contrary, it is a quite common story that appears every time we have to deal with new data (big or small, it does not matter). For example, at the early stage of genomic analysis, we have seen many dumb applications of standard statistical or machine learning techniques to, for example, microarray data. Of course, gene expressions do contain information, but standard machine learning or data mining techniques will not reveal any. Indeed, joint multiple-testing techniques and other ad hoc methods have been developed to assess, based on correct statistical grounds, the functional relationships between genes and pathologies. To summarize, when social scientists come to data or, even earlier, when social scientists approach data analysis, the main focus should not be the data per se or the model alone. The focus should be always on the correct statistical model *given* the data at hand. The blind application of machine learning, data mining techniques or other well-known methods may severely affect the results of the analysis no matter how good these techniques are for the specific task they were designed for.

Social science is not merely a set of tools but is mostly a way of looking at the world. That is social science is (or should also be) about how one thinks of problems and how he or she tries (eventually) to solve them. Such process, *from studying problems to solving them*, can explain why – at least in some contexts – "the influence of quantitative social science (including the related technologies, methodologies, and data) on the real world has been growing fast" (King 2014: 3).

Big Data are "today's data", and clearly every data are different beasts which require different methodologies. But these new data are more and more accessible and even if they require powerful backend and advanced statistical engines, these can be easily managed today from everyone's laptop even using open source tools. Computer science and information technology are certainly important as well as the connections with other disciplinary fields such as political communication and political science; but the general principle of looking into the data and thinking on the problem in a statistical way seems to be too seldom neglected.

Our results underpinned a theoretical framework that went in a different direction from the idea of building a representative sample of online population in order to carry out analysis at the individual level. To the contrary, the online crowds could be more useful to anticipate trends (and outcomes) when it is considered in the aggregate, as this can provide the finest description of the balance of power between different parties/candidates. What is more, our analysis has provided strong evidence in favor of mixing data from different sources in order to better understand public opinion dynamics. At the same time, however, our results confirmed that the method does matter and information available on social media should be analyzed using the proper techniques.

Of course, as with any other data, social media research will require replicability (King 1995) to gauge its viability as a source to study public opinion (see also Lacy et al. 2015 for a discussion on the best practices with respect to content analysis and social media data). Right now this is still (largely) lacking, due to the interplay of the massive amount of data analyzed (difficult to store and share on standard public platforms), their economic value and (sometimes) rather obscure and ad hoc algorithm adopted to produce the reported estimates (Sudulich et al. 2014). But at least with respect to this latter aspect, some steps have been taken lately in the right direction.

To conclude, it is time for the data to go back to social science or, better, for the social scientists (or data scientists, as it is now fancy to call people working on data) to start looking at the Big Data (with B and D capitalized) in an inclusive way. We consider this book as a contribution in this direction.

Notes

1 http://www.guardian.co.uk/technology/2011/mar/17/us-spy-operation-social-networks

2 http://www.digitalspy.com/music/news/a471915/justin-bieber-twitter-followers-50-percent-are-fake-says-report.html

3 http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html

4 This is a problem also for standard offline surveys, as the poll rates keep falling dramatically in recent years, thanks to mobile phones, caller identification and a rise in phone solicitation, while the difficulties of reaching many population segments still persist (Goidel 2011; Hillygus 2011; Tourangeau and Plewes 2013).

5 For example the percentage of population ages 55–64 increased on Twitter by 79% in the last year. See Global Web Index, "SOCIAL PLATFORMS GWI.8 UPDATE: Decline of Local Social Media Platforms." URL: https://www.globalwebindex.net/social-platforms-gwi-8-update-decline-of-local-social-media-platforms/

6 Notice that the number of observations was equal to 231 because in five cases we were not able to evaluate whether the prediction was made focusing on only one tweet per user as such information was not provided by the source of the forecasts.

7 Polling organizations also commonly use post-stratification weights to adjust the collected sample characteristics to match estimated population values (Little 1993).

8 Even when we exclude from the analysis the predictions based on SASA, which only produces an aggregate measure of sentiment and that, therefore, does not allow to distinguish the opinions of single users, the effect of *Weight* is still not statistically significant at the 95% level of confidence.

9 In other words, if for surveys of probability samples topic coverage follows naturally from population coverage, for social media analyses, topic coverage can, in principle, be achieved without population coverage. That is other mechanisms of information

propagation that are particular to the dynamics of social media may lead a collection of posts to accurately distill larger conversations in the full population despite the lack of population coverage among posters (Schober et al. 2016).

10 Interestingly, the contrast that we have often highlighted between methods that focus on the individual opinion of each single user versus methods that focus on the contrary directly on the aggregate opinion, reminds the long-lived (and heated) debate between those who see public opinion as the aggregate of individual opinions that can be recovered, for example, through surveys, and those who adopts an Enlightenment idea of public opinion that conceives it as the collective judgment arising during a public deliberation (Habermas 1980; Bourdieu 1979), i.e., as a "sharing opinion" rather than as a "aggregating opinion".

11 In a few cases we did not find any survey data and, therefore, it was impossible to compute the value of the variable *MAE Polls*. Accordingly, the number of observations dropped to 215.

12 Notice that if we add the variable *MAE Polls* in Models 1 and 2, Table 5.5 (Chapter 5) we find the same negative association between *MAE* and *MAE Polls*.

References

Ampofo, L., Anstead, N., and O'Loughlin, B. (2011) 'Trust, confidence, and credibility', *Information, Communication & Society*, 14(6): 850–871. doi: 10.1080/1369118X.2011.587882

Anderson, C. (2008) 'The end of theory: The data deluge makes the scientific method obsolete'. Available at: http://www.wired.com/science/discoveries/magazine/16–07/pb_theory

Bakker, T.P., and De Vreese, C.H. (2011) 'Good news for the future? Young people, internet use, and political participation', *Communication Research*, 20: 1–20.

Barberá, P. (2015) 'Who is the most conservative republican candidate for president?', *Washington Post*, June 16. Available at: https://www.washingtonpost.com/blogs/monkey-cage/wp/2015/06/16/who-is-the-most-conservative-republican-candidate-for-president/

Barberá, P., and Rivero, G. (2014) 'Understanding the political representativeness of Twitter users', *Social Science Computer Review*, 33(6): 712–729. doi: 0894439314558836

Best, S.J., and Krueger, B.S. (2005) 'Analyzing the representativeness of internet political participation', *Political Behavior*, 27(2): 183–216.

Bhutta, C.B. (2012) 'Not by the book: Facebook as a sampling frame', *Sociological Methods Research*, 41: 57–88.

Bourdieu, P. (1979) 'Public opinion does not exist', In Siegelaub, S. and Mattelart, A. (eds.), *Communication and class struggle*, 124–310. New York: International General/IMMRC.

Butler, D. (2013) 'When google got flu wrong: Us outbreak foxes a leading web-based method for tracking seasonal flu', *Nature*, February 13. Available at: http://www.nature.com/news/when–google–got–flu–wrong-1.12413

Ceron, A., Curini, L., and Iacus, S.M. (2016) 'iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content', forthcoming on *Information Sciences*, 367–368: 105–124.

Choy, M., Cheong, M., Ma, N., and Koo, P. (2011) 'A sentiment analysis of Singapore presidential election 2011 using Twitter data with census correction', *ArXiv*. Available at: http://arxiv.org/abs/1108.5520

Choy, M., Cheong, M., Ma, N., and Koo, P. (2012) 'US presidential election 2012 prediction using census corrected Twitter model', *ArXiv*. Available at: http://arxiv.org/abs/1211.0938

Conover, M., Ratkiewicz, J., Francisco, M.R., Gonçalves, B., Menczer, F., and Flammini, A. (2011) 'Political polarization on Twitter', *ICWSM*, 133: 89–96.

Cook, S., Conrad, C., Fowlkes, A., and Mohebbi, M. (2011) 'Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic', *PloS One*, 6(8): e23610.

Dalton, R. (2016) 'The potential of "big data" for the cross-national study of political behavior', *International Journal of Sociology*, 46: 1–13.

de Zúñiga, H.G., Nakwon, J., and Valenzuela, S. (2012) 'Social media use for news and individuals' social capital, civic engagement and political participation', *Journal of Computer-Mediated Communication*, 17(3): 319–336.

Diaz, F., Gamon, M., Hofman, J.M., and Kiciman, E. (2016) 'Online and social media data as an imperfect continuous panel survey', *PloS One*. doi: 10.1371/journal.pone.0145406

Dwi Prasetyo, N., and Hauff, C. (2015) 'Twitter-based election prediction in the developing world', *Proceedings of the 26th ACM Conference on Hypertext & Social Media* Guzelyurt, Cyprus, 1–4 September 2015.

Farrell, H., and Drezner, D.W. (2008) 'The power and politics of blogs', *Public Choice*, 134(1–2): 15–30.

Franch, F. (2013) '(Wisdom of the crowds)^2: 2010 UK election prediction with social media', *Journal of Information Technology & Politics*, 10(1): 57–71.

Franklin, M.N. (2002) 'The dynamics of electoral participation', In LeDuc, L., Niemi, R.G., and Norris, P. (eds.), *Comparing Democracies 2: New Challenges in the Study of Elections and Voting*. Thousand Oaks; London: Sage, 163.

Gayo-Avello, D. (2011) 'Don't turn social media into another "literary digest" poll', *Communications of the ACM*, 54(10): 121–128.

Gayo-Avello, D. (2013) 'A meta-analysis of state-of-the-art electoral prediction from Twitter data', *Social Science Computer Review*, 31(6): 649–679.

Gayo-Avello, D., Metaxas, P.T., and Mustafaraj, E. (2011) 'Limits of electoral predictions using Twitter', *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 17–21 July 2011..

Goidel, K. (2011) *Political Polling in the Digital Age: The Challenge of Measuring and Understanding Public Opinion.* New Orleans: LSU Press.

Habermas, J. (1980) *The Structural Transformation of the Public Sphere*. Cambridge, MA: The MIT Press.

Herbst, S. (2011) 'Un(Numbered) voices? Reconsidering the meaning of public opinion in a digital age', In Goidel, K. (ed.), *Political Polling in the Digital Age: The Challenge of Measuring and Understanding Public Opinion*, 85–98. Baton Rouge: Louisiana State University Press.

Hillygus, D.S. (2011) 'The evolution of election polling in the United States', *Public Opinion Quarterly,* 75(5): 962–981.

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

Huberty, M. (2015) 'Can we vote with our tweet? On the perennial difficulty of election forecasting with social media', *International Journal of Forecasting*, 31(3): 992–1007.

Iacus, S.M. (2014) 'Big data or big fail? the good, the bad and the ugly and the missing role of statistics', *Electronic Journal of Applied Statistical Analysis*, 5(11): 4–11.

King, G. (1995) 'Replication, replication', *PS: Political Science and Politics*, 28: 444–452.

King, G. (2014) 'Restructuring the social sciences: Reflections from Harvard's institute for quantitative social science', *PS: Political Science and Politics*, 47(1): 165–172.

Jensen, M.J., and Anstead, N. (2013) 'Psephological investigations: Tweets, votes, and unknown unknowns in the republican nomination process', *Policy & Internet*, 5(2): 161–182.

Jensen, M.J., Jorba, L., and Anduiza, E. (2012) ‘Introduction’, In Anduiza, E., Jensen, M.J., and Jorba, L. (eds.), *Digital Media and Political Engagement Worldwide: A Comparative Study*, 1–15. New York: Cambridge University Press.

Lacy, S., Watson, B.R., Riffe, D., and Lovejoy, J. (2015). ‘Issues and Best Practices in Content Analysis’, *Journalism & Mass Communication Quarterly*, 92(4): 791–811.

Lasorsa, D.L., Lewis, S.C., and Holton, A.E. (2012) ‘Normalizing Twitter: Journalism practice in an emerging communication space’, *Journalism Studies*, 13(1): 19–36.

Little, R. (1993) ‘Post-stratification: A modeler’s perspective’, *Journal of the American Statistical Association*, 88: 1001–1012.

Moy, P., and Murphy, J. (2016) ‘Problems and prospects in survey research’, *Journalism & Mass Communication Quarterly*, 93(1): 16–37.

Murphy, J.J. (2014) ‘Using respondent tweets to fill in survey gaps’, *Quirk’s Marketing Research Media*. Available at: http://www.quirks.com/articles/2014/20140125-1.aspx

Murphy, J.J., Keating, M.D., and Edgar, J. (2014) ‘Crowdsourcing in the cognitive interviewing process’, *Proceedings of the 2013 Federal Committee on Statistical Methodology Research Conference*, Washington, DC, 4–6 November 2013.

Nagler, J., and Tucker, J. (2015) ‘Drawing inferences and testing theories with big data’, *PS: Political Science and Politics* 48: 84–88.

Noble, I. (2003) ‘Human genome finally complete. BBC News’. Available at: http://news.bbc.co.uk/1/hi/sci/tech/2940601.stm

O’Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N.A. (2010) ‘From tweets to polls: Linking text sentiment to public opinion time series’, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, 23–26 May.

Plutzer, E. (2002) ‘Becoming and habitual voter: Inertia, resources, and growth in young adulthood’, *American Political Science Review*, 96(1): 41–56.

Rainie, L., Smith, A., Schlozman, K.L., Brady, H., and Verba, S. (2012) ‘Social media and political engagement’, *Pew Internet & American Life Project*. Available at: http://pewinternet.org/Reports/2012/Political-engagement.aspx

Rhodes, B.B., and Marks, E.L. (2011) ‘Using Facebook to locate sample members’, *Survey Practice*, 4(5). Available at: http://www.surveypractice.org/index.php/SurveyPractice/article/view/83/pdf

Sang, E., and Johan Bos, J. (2012) ‘Predicting the 2011 Dutch senate election results with Twitter’, *Proceedings of the Workshop on Semantic Analysis in Social Media*, Avignon, France, 23 April 2012.

Schober, M.F., Pasek, J., Guggenheim, L., Lampe, C., and Conrad, Frederick G. (2016) ‘Social media analyses for social measurement’, *Public Opinion Quarterly*, 80(1): 180–211. doi: 10.1093/poq/nfv048

Shi, L., Agarwal, N., Agrawal, A., Spoelstra, G., and Spolestra, J. (2012) ‘Predicting US primary elections with Twitter’, Unpublished manuscript. Available at: http://snap.stanford.edu/social2012/papers/shi.pdf

Sigelman, L., Roeder, P.W., Jewell, M.E., and Baer, M.A. (1985) ‘Voting and nonvoting: A multi-election perspective’, *American Journal of Political Science*, 29(4): 749–765.

Spierings, N., and Jacobs, K. (2014) ‘Getting personal? The impact of social media on preferential voting’, *Political Behavior*, 36(1): 215–234.

Sudulich, L., Wall, M., Gibson, R., Cantijoch, M., and Ward, S. (2014) ‘Introduction: The importance of method in the study of the ‘political internet’’, in Cantijoch, M., Gibson, R. and Ward, S. (eds.), *Analyzing Social Media Data and Web Networks: New Methods for Political Science*, 1–21. Basingstoke: Palgrave Macmillan.

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.

Tourangeau, R., and Plewes, T.J. (2013) *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: The National Academies Press.

Tsakalidis, A., Papadopoulos, S., Cristea, A., and Kompatsiaris, Y. (2015) ‘Predicting elections for multiple countries using Twitter and polls’, *IEEE Intelligent Systems*. doi: 10.1109/MIS.2015.17

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J.T., Nagler, J., and Tucker, J. (2013) ‘Social media and political communication: A survey of Twitter users during the 2013 Italian general election’, *Rivista Italiana di Scienza Politica*, 43: 325–355.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2014) ‘Forecasting elections with non-representative polls’, *International Journal of Forecasting*. Available at: http://www.sciencedirect.com/science/article/pii/S0169207014000879

Wei, L., and Hindman, D.B. (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.

Wilkerson, J., and Casas, A. (2017) ‘Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges’, *Annual Review of Political Science*, forthcoming.

Postscript (15 November 2016)

A lesson from the 2016 US presidential race: you cannot predict elections *without* social media data?

Soon after having submitted the final draft of our book, one of the most shocking elections in American political history (to say the least) took place. Given that we monitored the last month of the 2016 US presidential campaign day-by-day exactly as we did for the analogous race in 2012 (see Chapter 3), we are somehow compelled to add this short postscript.[1]

The triumph of Donald Trump the 8th of November, in one way or another, has been a huge surprise, but perhaps not equally a surprise for everyone. The results have shown a striking disconnection between the traditional media and polls on the one hand, and the electorate on the other. For weeks the media reported a virtually closed match with predictions based on surveys that gave Hillary Clinton a clear lead in the popular vote as well as in almost every swing state, quite often with a comfortable margin. Then came the election night, and we woke up with a very different reality than the one portrayed for weeks in the press.

True, media have misrepresented to a certain degree the interpretation given to polls data throughout the campaign,[2] but it is not heresy to argue that those same polls did not fare that well. Indeed, until the election night, several predictive models, mainly based on the analysis of survey data, assigned to Hillary Clinton a probability of victory that ranged from 72% – according to Nate Silver's popular blog *FiveThirtyEight* – to 85% (*New York Times*) to an astonishing 98%–99%, according for example to the *Huffington Post* and the Princeton Election Consortium.[3]

The prediction about the final votes share of the two main candidates was inaccurate too. Nate Silver attributed a margin of 3.6% to Hillary Clinton, while the estimates provided by Realclearpolitics.com, which – as we have already pointed out in Chapter 3 – computes the average value of survey polls, claimed that Clinton was leading by 3.3 points. Overall, the outcome of single surveys was even worse and polling companies working for ABC, CBS, FOX News, and *The Economist* gave to Hillary Clinton a 4-point advantage. Only the IBD survey, which considered Trump 2 points above Clinton, and the *Los Angeles Times*, which registered +3.2 for the Republican candidate, were suggesting that Trump could at least win the national vote.[4] Honestly, these predictions were not completely wrong given that the actual result falls within the margin of error of some of them. Nevertheless, survey data clearly failed to catch the strong signal coming from American voters. Why that? One possibility is related to what is called "confirmation bias", i.e., excluding the evidence that one does not like (or expect) and filtering in only what confirms a hypothesis. Possibly, however, the biggest reason goes back to what we already noticed in Chapter 4. Exactly as the "shy Tory" thesis in the UK elections has been set forward as an explanation for the mismatch between the prior forecasts and the election outcome, the same could be argued for a "shy Trump voter" effect. Accordingly, Trump supporters could have refrained from expressing their voting choice in survey polls, either because they were distrustful of institutions, including polling companies, or because after the scandals reported in the media involving the Republican candidate, they felt uncomfortable expressing their actual voting behavior.[5] The "confirmation bias effect" combined with the "shy Trump voter effect" to generate a "perfect storm" that produced a very weak performance of survey polls (without mentioning the extremely poor results of the exit polls[6]).

However, as we have argued throughout the present book, people could feel more free to express online their personal views without being affected by conformism and social desirability. And this is what, after all, happened, at least according to our analysis. To monitor the evolution of American public opinion, we followed our usual approach illustrated at length in the previous pages, focusing on comments coming from the United States that were discussing explicitly about the two candidates on Twitter. (Overall, we have analyzed around 32.5 million posts on a daily basis, excluding the huge peaks during the presidential debates). However, we also decided to apply what is discussed in Chapter 5; that is, we integrated survey data with the information on the mood of social media analyzed by means of iSA as follows: from 19th of September till the 2nd of October, we ran an econometric model in which we used the results of our sentiment analysis to predict on a daily basis the estimates of national survey data as published on Real Clear Politics web-page. It turned out that by using just the percentage of negative comments toward Donald Trump and the online voting intentions expressed in support of Hillary Clinton we were able to explain 97% of the overall trend in the survey. Then, beginning on the 3rd of October, we relied on the just mentioned two social-media indicators, conveniently weighted according to the analysis just illustrated, to produce our estimates of the electoral forecast.

According to our analysis, at the beginning of October Donald Trump appeared to overtake Hillary Clinton in preferences (see Figure PS.1). Then, Trump faced difficulty when attacked on tax evasion, and even more after the video published by the *Washington Post* in which Trump denigrated the image of women. If the sex video scandal brought down Trump in early October, the problems involving Clinton's campaign (highlighted by WikiLeaks) and the desire to "drain the swamp", favored Trump's recovery. Even discounting for the propaganda bots used by Trump staff (estimated to be about 20%–30% of the total volume[7]), the sentiment toward Trump pushed upwards, and in the second half of October, although polls

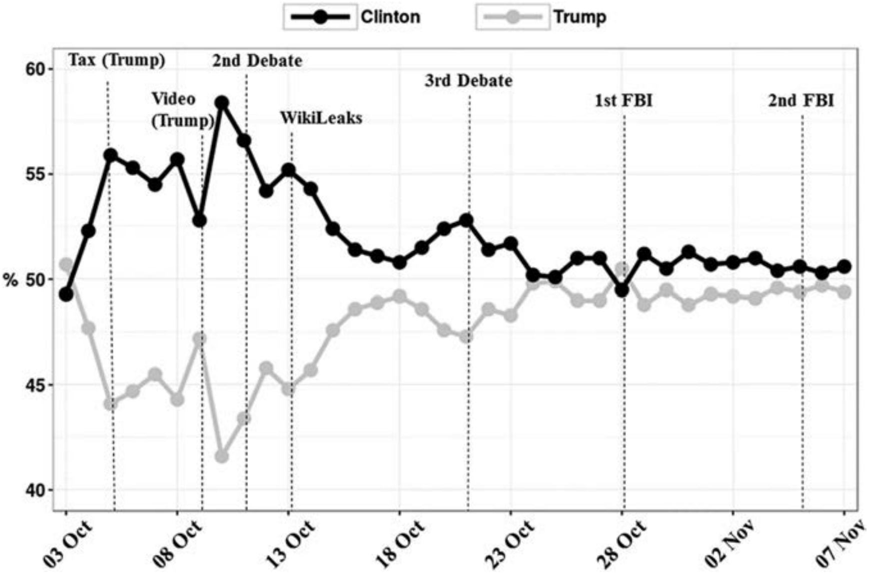

Figure PS.1 2016 US presidential election: daily social voting intentions according to iSA

showed Clinton largely ahead, the Republican candidate was sharply rising, so much so that according to our estimates, on the day in which the FBI director "re-opened" the email case against Hillary Clinton, Trump climbed over Clinton in the popular vote. The race then remained narrow, a sort of "too close to call" contest, with Clinton oscillating around the only vantage point (+1.2 our final forecast). Hillary Clinton indeed won the popular vote by 0.6 points (at least according to the updated vote-counting procedure reported as in 15th November), far away from the 3–4 margin points that were predicted on the eve by pollsters.

But, as we all know, in American elections what really matter are the swing states. Among the 14 swing states considered, on 10 occasions our forecast proved to be correct (Florida, Ohio, North Carolina, Georgia, Arizona, Virginia, Colorado, Nevada, New Hampshire, Minnesota). Here, too, our data was able to capture some important signals that escaped other analysts. On Twitter, for example, the Trump victory in Ohio and Florida was never questioned, even in mid-October, although other analysts regarded Florida as Clinton's territory. Similarly, the sentiment predicted a fairly easy victory for Hillary Clinton in Nevada and Colorado, while the surveys considered the two were very much in the balance. And while no one cared about Pennsylvania, we regarded this as a key state, as it was indeed. Five days before the vote, our estimate recorded an advantage of Trump in Pennsylvania, so that on that day, our forecast predicted Trump as the next president of the United States.[8]

In a nutshell, we forecasted a more open game than others did, giving Trump a good percentage of success rate, predicting the victory of the Republican candidate in some key states and coming very close to the real gap between the two candidates in the popular vote. Contrary to media enthusiasm for the oncoming Hillary Clinton victory, our model suggested a higher level of uncertainty: if throwing 10 times the "dice", in 4.1 circumstances, we would in fact predict a victory for Trump.[9] And in the end, Donald Trump won.

Our estimates proved therefore to be more accurate than survey polls alone, suggesting that combining two sources of data on public opinion can produce gains in terms of accuracy, exactly as discussed in Chapter 5. In fact, survey data can mitigate the bias of social media, while social media can attenuate the bias of survey polls. In this respect, the appropriate mixture of multiple sources of data seems nowadays the best way to really get the whole picture, if our aim is to understand the behavior of a volatile public opinion.

The 2016 American elections confirm once again the ability of social networks to grasp in advance current trends in public opinion and in society at large. Obviously, as in all disciplines, the right tools are needed, and this book is about methods after all. Still, we cannot deny that public opinion has profoundly changed. The way to measure it must change as well: there is no longer the option to avoid listening to social media information. This is the path for better nowcasting (and forecasting) politics.

Notes

[1] http://sentimeter.corriere.it/2016/10/17/presidenziali-americane-2016-secondo-la-rete/

[2] http://www.realclearpolitics.com/articles/2016/11/12/it_wasnt_the_polls_that_missed_it_was_the_pundits_132333.html

[3] http://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html

[4] http://www.realclearpolitics.com/epolls/2016/president/us/general_election_trump_vs_clinton_vs_johnson_vs_stein-5952.html

[5] This could have been true especially for women backing Donald Trump. See http://fivethirtyeight.com/features/the-polls-missed-trump-we-asked-pollsters-why/?ex_cid=2016-forecast

[6] http://www.slate.com/votecastr_election_day_turnout_tracker.html

[7] http://firstmonday.org/ojs/index.php/fm/article/view/7090/5653

[8] http://sentimeter.corriere.it/2016/11/03/se-si-votasse-oggi-il-presidente-sarebbe-trump-almeno-per-la-rete/

[9] http://www.corriere.it/elezioni-presidenziali-usa-2016/notizie/trump-nuovo-presidente-usa-segnali-sottovalutati-due-lezioni-imparare-sull-opinione-pubblica-473743a8-a72e-11e6–8208–49eea13f646a.shtml

Index

Aarts, K. 98
administrative data 1
Affordable Care Act 78
agenda-setting power 7–8
agglomerative methods 47–8
Agrawal, M. 12
Alfano, Angelino 116, 122–5
American Idol 14
Amrit, C. 98
Antweiler, W. 13
Arab Spring 9, 13
Augustine, C.B. 14
Avery, J.M. 6

backlash effect 121
Barclay, F.P. 19
Barnes, M.D. 14
Barrett, P. 19
Basilaia, E. 8
Benghazi-gate 81
Bennett, W.L. 9
Berlusconi, Silvio 107, 110, 116–17, 122–5, 133
Bermingham, A. 22
Bersani, Pier Luigi 39, 87–94, 106, 110, 116, 123, 133, 134
Big Data 1–2, 39, 154–5; results weighted against individual data 155–61
Bin Laden, Osama 13, 80
Bloomberg, Michael 81–2
Bos, J. 21, 22
Bossi, Umberto 94–9, 133
Burnap, P. 12
Burton, S.H. 14

Cambraia, C.A. 19
Cameron, M.P. 19
campaigning, electoral 105; effectiveness of positive and negative Twitter 113–18; empirical analysis and results 118–22; online voting intentions in 2013 Italian election and 105–13; relationship between leaders sentiment and voting intentions and 122–6
Campante, F. 6
Cancellieri, Annamaria 97, 99
Canova, L. 12
Casini, Pier Ferdinando 116, 123, 133, 134
Ceron, A. 8
Chelala, P. 21
Chiao, C. 85
Christie, Chris 81–2
CIA (Central Intelligence Agency) 13
Civati, Giuseppe 95, 96–7, 98–9
classification methods 47–52
clicktivism 9
Clinton, Hillary 171–4
cluster analysis 47–8
cluster ensembles 48
Coletto, M. 97
collective action 9–10
computational approach to social media-based electoral forecasts 16–20
Comte, F. 85
confirmation bias 172
Conway, B.A. 7
Corriere della Sera 95
Coskun, M. 13
Cristea, A. 161
Cristianini, N. 14
criticisms of social media-based electoral forecasts 22–4
Cunha, E. 19
Cunningham, J.A. 14
Cuperlo, Gianni 95, 96–7, 99
Curini, L. 8, 12

Dassen, A. 98
data mining 47
Dave, K. 39

de Jong, F. 85
de Voogd, L. 21
Di Pietro, Antonio 106, 123
dissimilarity 47
dissociative methods 47
Dotterweich, L. 6
Downs 142
Durahim, A.O. 13
Durante, R. 6
dynamic multi-topic models 49
Dziurzynski, L. 12

Eastwood, Clint 79
e-campaigning 10–11
Eichstaedt, J.C. 12
emotions and well-being 12–13
endorsement data 19
estimation of policy positions using SNS 11–12

Facebook 38; Arab Spring and 9; Big Data revolution and 2; electoral forecasts and 15; endorsement data 19; growth of 5; sentiment analysis 21; voting behavior declarations on 18
Falck, O. 6
Financial Times 154
Fini, Gianfranco 123, 133
first-level agenda-setting 7–9
forecasts, electoral 132; best practices 154–5; comparing the approval of Italian leaders in 2011 and 132–5; computational approach to social media-based 16–20; discussion of SASA method for 149–51; limits and criticism of 22–4; made through SASA 68–71; meta-analysis of existing studies 143–8; sentiment analysis and machine learning applied to 20–2; 2015 Veneto regional election 135–8; 2012 French legislative elections 138–43; 2012 US presidential election 76–85; use of social media in 14–16; weighting results with individual data in 155–61; *see also* nowcasts, electoral
France: legislative elections, 2012 138–43; presidential election, 2012 71–5
Franch, F. 19
Frank, M.Z. 13
functional words 40

Giannino, Oscar 112
Giraud-Carrier, C. 14
GitHub 54, 59
Gold, R. 6
Gonçalves, M.A. 19
Google 13; Flu Trends 14, 154; Insights 8; Trends 8
Grillo, Beppe 106, 112, 116, 123, 124–5
Grimmer, J. 48
Guardian, The 154
Guggenheim, L. 8

Hale, K. 6
Hansen, B. 14
Hansen, H.M. 14
Hanson, C.L. 14
hashtags 69–70; *see also* Twitter
Hayashi-Yoshida asynchronous covariance estimator 86
Heblich, S. 6
Hierarchical methods 48
Hoffmann, M. 85–6
Hollande, François 71–5
Hopkins, D. 39, 51, 52
Hosch-Dayican, B. 98
Howard, P.N. 9
human and automatic tagging 46–7
Human Genome Project 154
Hung, K. 85
Hurricane Sandy 81–2
Hussain, M.M. 9

Iacus, S.M. 8, 12
Ingroia, Antonio 106, 112, 116
iSA (integrated Sentiment Analysis) 39–40, 52–4; Donald Trump data example 59–65; online voting intentions and 107–13; United States presidential election, 2016 and 172–3
iSAX package 54–8
Italy: comparing approval of leaders in 2011 and election forecasts 132–5; effectiveness of positive and negative Twitter-campaigning in 2013 election in 113–18; online voting intentions in 2013 election 105–13; party leader primaries, 2013 94–9; primary election, 2012 87–94; 2015 Veneto regional election 135–8

Jha, S. 12
Jiang, Y. 7
Johnson, T.J. 6
Jungherr, A. 8

Kang, J. 85
Kaye, B.K. 6
Keim, D.A. 48
Kenski, K. 7
Kern, M.L. 12
Kim, A. 14

King, G. 39, 48, 51, 52
Kompatsiaris, Y. 161
Kroutil, L.A. 14

Lakshmikanth, S.K. 12
Lampos, V. 14
Lawrence, S. 39
LDA (latent Dirichlet allocation) method 48–9
lead-lag estimation 85
Lee, C. 85
Lee, S. 85
Lindsay, R. 21, 22
Lucas, R.E. 12
Lucchese, C. 97

machine learning 21–2, 49–52
Mcneal, R. 6
MacWilliams, M.C. 11
Magno, G. 19
Maroni, Roberto 116, 123
mean absolute error (MAE) 23
Meraz, S. 7
Mo Jang, S. 8
Monti, Mario 106–10, 116–17, 123
Moretti, Alessandra 136–8
Mourdock, Richard 81
Murphy, J.J. 14, 161

negative campaigning, effectiveness of 113–18; empirical analysis and results 118–22
Newman, R.W. 8
New York Times 13
Nijman, T. 85
noise 50–1; iSAX package and 54–8
nowcasts, electoral 68, 70–1; Hollande versus Sarkozy in 2012 French presidential election 71–5; *see also* forecasts, electoral

Obama, Barack 20, 21, 38; Twitter analysis using SASA approach and 76–85
Obamacare 78
Occupy Wall Street movement 9
old versus new media 7–9
Open Source Indicators (OSI) 13
opinion mining 39
opinion polls 161–4
organic data 1–2
Orlando, S. 97

Papadopoulos, S. 161
Park, G.J. 12
Parmelee, J.H. 7
Pennock, D.M. 39
Perego, R. 97
Pew Research Center 18
Philipp, G.S. 18
Pichandy, C. 19
Pittella, Gianni 95
Polgreen, P.M. 14
politics: flourishing relationship between society, social media and 5–14; impact of the Web on 5–7
positive campaigning, effectiveness of 113–18; empirical analysis and results 118–22
Prodi, Romano 110
public policy 9–10

quantitative and qualitative text analysis 38–9

ReadMe algorithm 52
Realclearpolitics.com 82
Recorded Future 13
Renault, E. 85
Renzi, Matteo 38–9, 87–94, 95, 96, 97, 98–9, 110
Richards, A.K. 14
Rich Site Summary (RSS) feeds 16
Robert, C. 85
Romney, Mitt 20, 38; Twitter analysis using SASA approach and 76–85
Rosenbaum, M. 85–6
R statistical environment 54, 59–65

Sage, A.J. 14
Salvini, Matteo 94–9
Sarkozy, Nicolas 71–5
SASA approach 68–71, 149–51; France presidential election, 2012 71–5; Italy primary elections 87–99; online voting intentions and 111; United States presidential election, 2012 76–87
Scharkow, M. 8
Schwartz, H.A. 12
Schwarzer, S. 21
scoring techniques in text analysis 44–6
second-level agenda-setting 8–9
Segerberg, A. 9
Segre, A.M. 14
Self-Organizing Maps (SOM) 48
self-selection bias 24
Seligman, M.E. 12
sentiment analysis (SA) 20–2, 38, 164–6; Donald Trump data example 59–65; opinion mining and 39; principles of 39–41; quantitative and qualitative analysis of texts in 38–9; relationship to opinion polls 161–4; stemming in 41–4; *see also* text analysis

Servizio Pubblico 124
Shaw, D.L. 8
Signorini, A. 14
Silver, Nate 171
slacktivism 9
Smeaton, A. 22
Sobbrio, F. 6
social desirability 111, 127, 164, 172
social media and social network sites (SNS): agenda-setting power 7; analysis relationship to opinion polls 161–4; best practices for using data from 154–5; collective action and public policy influence 9–10; compared to traditional media 7–9; data 2; e-campaigning using 10–11; economic, social and political forecasts using 13–14; electoral forecasts and 14–16; emotions and well-being measured using 12–13; estimating policy positions using 11–12; flourishing relationship between politics, society and 5–14; online voting intentions in 2013 Italian election and 105–13; *see also* Twitter
spatial theory of voting 114
Sprenger, T.O. 18
stemming 41–4
Stephens, M. 12
Stewardson, B. 19
Sudhakaran, S. 19
Surveys Voting Intentions 125

text analysis: classification methods 47–52; fundamental principles of 39–41; making text digestible to statistical models for 41–4; opinion discovery and 38; pros and cons of human and automatic tagging in 46–7; quantitative and qualitative 38–9; scoring techniques 44–6
text mining 47
Tjong Kim Sang, E. 21, 22
topic models 48–9
traditional sentiment analysis 21
training set 41
transaction data 2
Trump, Donald 59–65; Twitter analysis using iSA approach and 172–4
Tsakalidis, A. 161
Tufekci, Z. 9
Tumasjan, A. 18
Twindex 84
Twitter 38; agenda-setting power 7, 8, 9; analysis of mentions of parties and candidates on 15; Big Data revolution and 2; drug abuse analysis 14; effectiveness of positive and negative campaigning on 113–18; endorsement data 19; estimating policy positions using 11; growth of 5; “hate map” 12; nowcasting using 69–70, 71–5; reciprocal influence between new and old media and 8; Recorded Future 13; SASA approach and 68–70; sentiment analysis 20, 21; social media crowd and 18; voting behavior declarations on 18; *see also* forecasts, electoral; social media and social network sites (SNS)

Ungar, L.H. 12
United States presidential election, 2012: surveys versus sentiment in 85–7; Twitter analysis using SASA approach and 76–85
United States presidential election, 2016: Twitter analysis using iSA approach and 172–4

Vargo, C.J. 8
Vendola, Nichi 116, 123, 133, 134
Venkat, A. 19
Véronis, J. 16
Virgilio, A. 19
Vogelgesang, J. 8
volume data 19
voting intentions: online 105–13; relationship between leaders sentiment and 122–6

Wallsten, K. 7, 8
Wang, D. 7
Welpe, I.M. 18
West, J.H. 14
Wikipedia 15
Williams, M.L. 12
Wilson, C. 9
Woodly, D. 7
Wordfish 44–6
Wordscores 46

Yoshida, N. 85–6
Young Bae, S. 8
YouTube 8

Zaia, Luca 136–8