

Can electoral popularity be predicted using socially generated big data?

Taha Yasseri
Oxford Internet Institute, University of Oxford, 1 St Giles', OX1 3JS Oxford, UK

Jonathan Bright
Oxford Internet Institute, University of Oxford, 1 St Giles', OX1 3JS Oxford, UK

Abstract

Today, our more-than-ever digital lives leave significant footprints in cyberspace. Large scale collections of these socially generated footprints, often known as big data, could help us to re-investigate different aspects of our social collective behaviour in a quantitative framework. In this contribution we discuss one such possibility: the monitoring and predicting of popularity dynamics of candidates and parties through the analysis of socially generated transactional data on the web during electoral campaigns. We focus on two data streams, Wikipedia page views and Google search queries, and discuss issues related to data collection, data cleaning, data analysis as well as the expressiveness and representativeness of the data; we also present popularity dynamics from real case examples of recent elections in three different countries.

Keywords: E.m [Data:Miscellaneous]; J.4.c [Applications: Social and Behavioral Sciences: Sociology]; J.8.o [Applications: Internet Applications:Traffic analysis]; J.2.h [Applications: Physical Sciences and Engineering: Mathematics and statistics]; J.8.j [Applications: Internet Applications: Libraries/information repositories/publishing]; K.4.3.b [Computing Milieux: Computers and Society: Computer-supported collaborative work]

MS-ID: / ()	taha.yasseri@oii.ox.ac.uk	December 11, 2013
----------------	---------------------------	-------------------

1 Introduction

Increasing use of the internet, and especially the rise of social media, has generated vast quantities of data on human behaviour, significant portions of which are also readily available to researchers. The potential of these data has not gone unnoticed: in just a few years use of social media data in particular has moved from minority pursuit to become almost mainstream in the social sciences, leading to the emergence of a new paradigm of “computational Social Science” [20, 21]. One of the most intriguing possibilities raised by the emergence of social media data is that it could be used to supplement (or even eventually replace) traditional methods for public opinion polling, especially the sample survey, because social media data offer considerable advantages in comparison with surveys in terms of the speed with which they can be acquired and the cost of collection. The selection bias in social media is clear: not everyone uses it, and people who do are not randomly distributed throughout the population [13]. Yet the hope has frequently been expressed that the sheer quantity of social media users may start to compensate for this (around 50% of the UK’s population are thought to have a Facebook account, for instance) and that might eventually replace the “sample-based surveys” with the “whole population data”.

The potential applications of “social polls” are wide ranging, however probably the most frequently explored avenue of research has been the use of social media data for electoral prediction (see e.g. [11, 16, 10, 12]). This is because the outcomes of elections are interesting in and of themselves, but also because it is a subject where a huge amount of validation data exists, coming from both the more traditional opinion polling which social media data might hope to replace, and the results of the election itself. Such social polling is typically based on one of two main methodologies: either offering a simple count of all tweets mentioning a given candidate, or using various techniques developed for analysing the sentiment expressed in them as a measure of people’s opinion on a given candidate.

Despite initial enthusiasm, and in contrast to the cases of predicting arrival of earthquake waves [18] and traffic jams [19], most of the recent research on using Twitter for electoral prediction has been relatively negative, with many researchers reporting weak correlations with actual electoral outcomes, difficulty duplicating other positive research, or rates of successful prediction that could easily have come about by chance (see *inter alia* [15, 14]). Most problematically, results have often exhibited specific biases either for or against individual political candidates, with minority parties often systematically overstated (see [10]), whilst major conservative candidates are often understated [14].

A variety of potential reasons have been put forward for these problems. The most obvious is that the self-

selection problem of social media cannot in fact easily be overcome with a larger sample size. Self-selection also operates when users decide what to post: even if large amounts of the population have created social media accounts, the amount which use them to express political opinions is much more limited. The nature of social media also means that opinions which are expressed are those heard by friends, family, work colleagues and other social connections: which might compel people to moderate their opinions or keep quiet if they support particular types of political party. Furthermore, many researchers have observed the difficulty of reliable sentiment analysis of political tweets, both because of the small amount of information contained in any given tweet and because of the nuances of political language where many opinions are expressed through irony or sarcasm [14]. Finally, as social media have started to take on a prominent position in media landscape (with trending topics now frequently a basis for news stories), political candidates have also increasingly started to intervene actively in social media, which has the potential for biasing results [1].

1.1 Google Trends and Wikipedia Page Views: Predicting the Present

While social media data are probably the most used of the new data sources which have been generated by the internet, significant interest has also arisen surrounding the use of informational search data present in websites such as Google Trends or Wikipedia, which is generated when someone either conducts a web search for a particular topic or accesses a particular page on Wikipedia. While not typically regarded as social “media”, search data is nevertheless socially generated in that it relies on people entering individual search queries. Having clear information on what people are looking for when they are looking for it provides a number of opportunities to “predict the present”: to gain a kind of real time awareness of current behaviour patterns. Such data have already been used to successfully predict a wide variety of phenomena both in short and long terms, from car and house prices to trends in flu outbreaks or unemployment [7, 6, 8, 9] using web search data, as well as movies box office revenues using Wikipedia page view statistics [17].

Information seeking data offers significant theoretical advantages when compared to social media data in terms of its use for prediction. Whereas the automatic interpretation of the meaning of a tweet can be riddled with complexity, the interpretation of the meaning of a search or the access of a page in Wikipedia is much more straightforward: the user is interested in information on the topic in question. Furthermore, the penetration of search especially is far greater than many social media platforms, especially Twitter (approximately 60% of internet users use search engines [5]). However, such data has rarely been applied to the task of election predic-

tion. The main reason for this is simple: “queries are not amenable to sentiment analysis” [4]. When entering a search query people express what they are looking for, but not their opinion about the subject: indeed, given they are searching for information, it seems reasonable to assume that this opinion is not fully formed. Despite this problem, in this paper we argue that there is significant “sentiment” data implied in information seeking behaviour. In the same way that searches for car models or stock types imply people are already considering a purchase, we expect that searches for political candidates imply that people are already considering a vote.

The problem for the purposes of prediction is that the relationship between search traffic and actual outcomes is unlikely to be straightforward. In fact, in one of the few studies that have attempted to apply information seeking data to elections [2], simply using search volume in the days prior to the election is an extremely poor prediction technique. Rather, we argue, there are a number of intervening variables which may affect how people look for information on politics, and thus need to be taken into account. One obvious first factor would be whether the political system encourages focus on parties or individuals (which may itself emerge through different modes of democratic organisation, e.g. presidentialism vs. parliamentarism), something which is likely to affect the search terms people enter. Also worth considering is the amount of potential candidates on the political scene, with elections full of new faces likely to generate more searching than contests between familiar candidates. Finally, there is the extent to which the existing incumbent is popular: as people are more likely to be informed on what the current power holder’s views are, they are less likely to search for them.

Within the context of this paper, we seek to explore some of these questions by looking at correlations between search engine data, Wikipedia usage patterns and recent election results in three different countries: the UK, Germany and Iran. These countries were selected in order to provide a diverse range of political contexts (with elections in Iran and the UK where a new candidate was voted in and one in Germany where a popular incumbent was returned), electoral systems (from Iran’s presidential system to the parliamentary ones operated in the UK and Germany) and party landscapes (with a very stable system in the UK contrasted to Iran and Germany where new actors are emerging).

2 Data Collection

For our analysis, we collected data from both Google Trends and Wikipedia for the last election in each of our countries of interest (2013 in the case of Iran and Germany, 2010 in the UK). Our trends data is based on the amount of searches for either a given party or politician coming from our specific country of interest

(search terms were entered in the native language and script of that country).

Our Wikipedia data is extracted from the page view statistics section of the Wikimedia Downloads site (<http://dumps.wikimedia.org/other/pagecounts-raw>) through the web-based interface of “Wikipedia article traffic statistics” (<http://stats.grok.se>); again, for Wikipedia we focus on language specific terms appropriate to the country of interest.

3 Results

We will begin with a discussion of the Iranian election of the 14th of June 2013. Iran operates a presidential system, where individual candidates are far more important than political parties. The presidency goes to the candidate who gains more than 50% of the vote, with a runoff in case no candidate is able to in the first round. The election of 2013 was an unusual one: it lacked an incumbent candidate (with former president Mahmoud Ahmadinejad standing down after fulfilling the maximum two terms in office), and was won convincingly by Hassan Rouhani in the first round, a candidate who was perceived as an outsider just a month before the election. Figure 1 shows patterns in Wikipedia page views and Google Search volume for the Iranian election, whilst the final results can be seen in Table 1. Several patterns are immediately apparent. First, both Google and Wikipedia call the winner of the election correctly, and also pick up on the large absolute disparity between Rouhani and the other candidates. They are both also sensitive to the very late development of Rouhani as a candidate (though Wikipedia also shows a spike in May). This comes as a very interesting result as none of the official polls have predicted the victory of any candidate in the first round of the election (with the highest 42% for Rouhani) [3]. However, neither Google nor Wikipedia correctly identify second place.

We will now move on to the German election of the 22nd of September 2013. Germany operates as a federal parliamentary republic, with power divided between the German parliament (“Bundestag”) and the body which represents Germany’s regions (“Bundesrat”). This election in particular was for the Bundestag, which itself has responsibility for electing Germany’s Chancellor, its most powerful political office. Germany’s system is based strongly around parties: a majority vote is required to elect the Chancellor, which is usually based on a coalition between one or more parties. In this particular election, the winning Christian Democrat party (CDU/CSU) increased its vote share for the second successive election, confirming its place as a highly popular incumbent party. However its coalition partner from the 2009 elections (the FDP) lost a lot of ground, failing to win any seats, meaning that the absolute outcome of the election is still uncertain while negotiations proceed

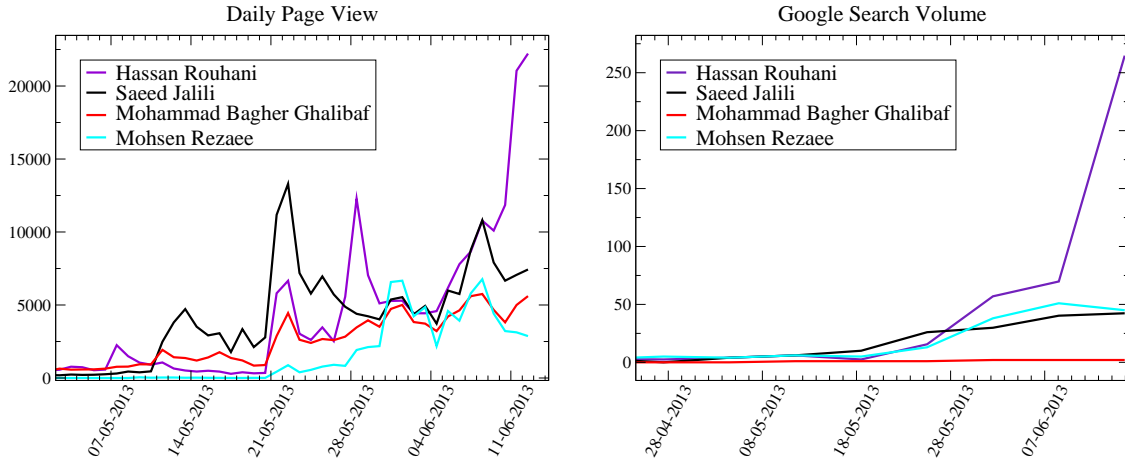


Figure 1: The time evolution of Wikipedia page views and Google search volume for the four leading candidates of Iranian presidential election of 14 June 2013 are shown in the left and right diagrams respectively.

Candidate	Popular Vote	Percentage
Hassan Rouhani	18,613,329	50.88
Mohammad Bagher Ghalibaf	6,077,292	16.46
Saeed Jalili	4,168,946	11.31
Mohsen Rezaee	3,884,412	10.55

Table 1: Final results of the Iranian presidential election, 14 June 2013.

over the potential formation of a different coalition.

The results of the election are shown in Table 2, whilst the data extracted from Wikipedia and Google are shown in Figure 2. The results show an interesting contrast to the Iranian election. Google predicts correctly both the winner of the election and second place (if we look at the date of the election), and is also approximately right about the distance between the two parties. It radically overstates the position of the FDP however. Wikipedia, by contrast, does not predict anything accurately, overstating to a large extent the position of Alternative for Germany, a radical anti-Euro party which was recently formed. This chimes with earlier work by Jungherr [10] who found that Twitter overstated to a large extent the position of the Pirate Party in the 2009 German election.

We will now look finally at the results of the 2010 UK election. The UK also operates a parliamentary system, though unlike Germany does not have a separate regional body. Rather, power is concentrated on one legislative body (the House of Commons), with a secondary unelected body (the House of Lords) providing some checks and balances. The history of the UK has been dominated by single party government, as the voting system there favours the emergence of a small group of very large parties. Hence even though in theory parliament and hence parties elect the prime minister, in practice the individual personalities of leaders have come to be seen as just important as party identity. For

this reason in the UK we look at both individuals and parties.

Figure 3 shows results from Wikipedia and Google for the UK election, whilst Table 3 reports the actual results. A variety of findings are worth noting here. Firstly, on Google, parties were universally more searched for than politicians, however the party data itself did not offer a useful predictor of the election results, considerably overstating the position of the Liberal Democrats, the UK's third largest party (though this party did improve considerably on its 2005 result). The individual politician data did, by contrast, place all the winning parties in correct order, though the difference between Conservative candidate David Cameron and Labour candidate Gordon Brown was marginal. In Wikipedia, by contrast, individual politicians were much more viewed than parties. Both the politician and party data offers a correct placement of all four parties, though the differences between them are microscopic.

4 Discussion and Conclusion

There are several broad conclusions we would like to draw from this data. It is clear first and foremost that online information seeking forms a part of contemporary elections: all three of the countries under study showed significant increases in traffic in the days leading up to an election. However it is also clear that patterns dif-

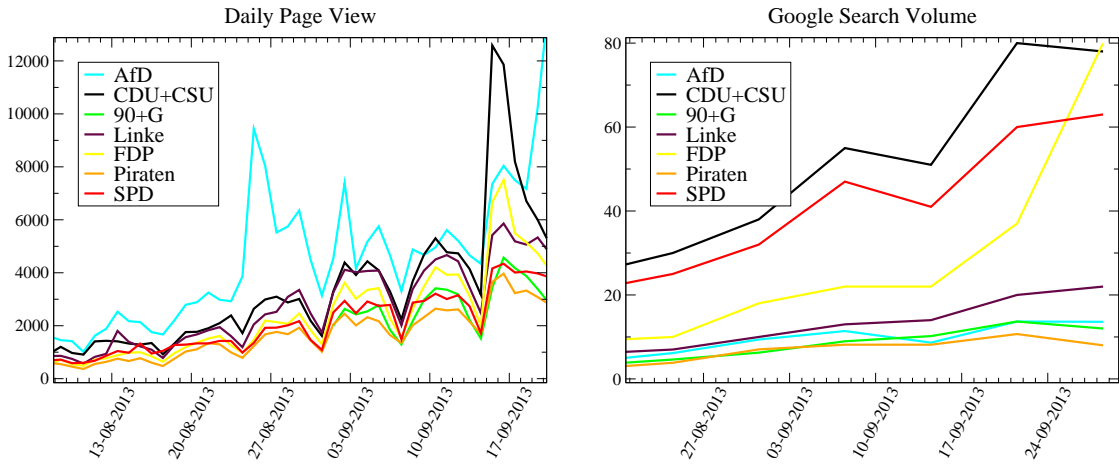


Figure 2: The time evolution of Wikipedia page views and Google search volume for the 7 leading German parties during the 22 September 2013 parlimentary election campaign are shown in the left and right diagrams respectively.

Party	Popular Vote	Percentage
Christian Democratic Union (CDU)	14,921,877	34.1
Social Democratic Party (SPD)	11,252,215	25.7
The Left (DIE LINKE)	3,755,699	8.6
Alliance '90/The Greens (GRÜNE)	3,694,057	8.4
Christian Social Union of Bavaria (CSU)	243,569	7.4
Free Democratic Party (FDP)	2,083,533	4.8
Alternative for Germany (AfD)	2,056,985	4.7
Pirate Party (PIRATEN)	959,177	2.2

Table 2: Final results of the German federal election, 22 September 2013.

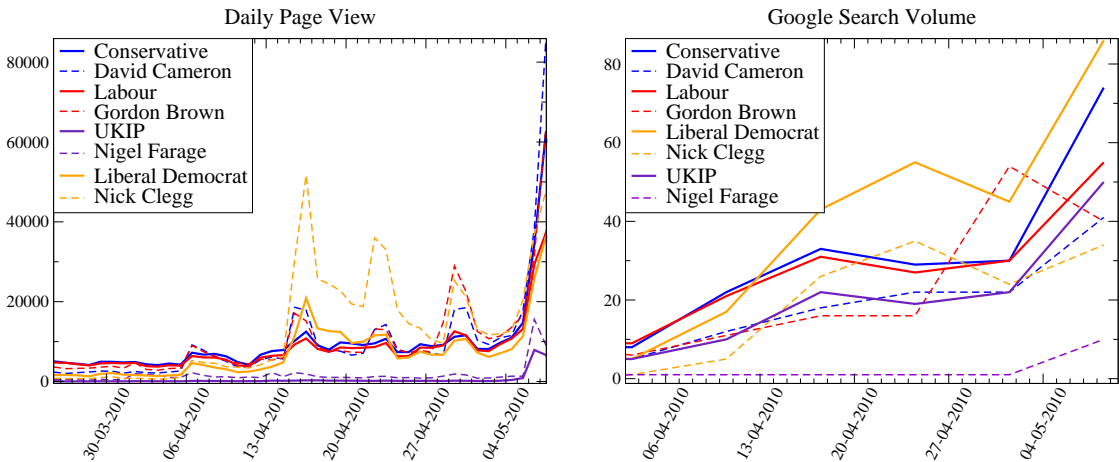


Figure 3: The time evolution of Wikipedia page views and Google search volume for the four leading parties and their leaders during the 6 May 2010 UK general election campaign are shown in the left and right diagrams respectively.

Party	Popular Vote	Percentage
Conservative	10,703,654	36.1
Labour	8,606,517	29.0
Liberal Democrat	6,836,248	23.0
UKIP	919,471	3.1

Table 3: Final results of the United Kingdom general election, 6 May 2010.

fer in the context of different elections, and that people do not simply search in the same proportions that they vote. Even the overall patterns show dissimilarities, while German data shows a clear weekly pattern, with the minimum of volumes during weekends, such patterns are absent in other two countries.

We highlight several key factors here. Firstly, data based on individual politicians proved more reliable than data based on parties: both Wikipedia and Google predicted the winners of the Iranian and UK elections when using individual politicians as search terms. We expect this is because there is a greater variety of ways in which people can search for information on a political party than there is an individual (they could, for example, use an abbreviation, or search for “Labour Party” rather than “Labour”).

Secondly, it is clear that information seeking data reacts quickly to the emergence of new “insurgent” candidates, such as Hassan Rouhani or the AfD. However, supporting previous work, it may also overstate them (the high volumes for the Liberal Democrats in the UK can also be read in this light). For this reason, it may be useful for social predictions to look for multiple different information sources. The AfD, for example, performed well in Wikipedia but poorly on Google, whilst the reverse was true for the Liberal Democrats. Rouhani, by contrast, performed well on both platforms. Finally, it seems that information seeking data is at its least effective when predicting the decline of a previously popular party. The FDP provides the example here: there is little to suggest in either Google or Wikipedia that it was about to suffer the reverse it did.

In conclusion, we argue that there is significant potential in information seeking data for both enhancing our knowledge of how contemporary politics work and predicting the outcome of future elections, especially the high pace and low cost of this approach compared to classic polling. However much work remains to be done in establishing the conditions under which such prediction will be successful. In our view, this will depend on elaborating more fully a theory of how people seek information on politics, and how different electoral circumstances change this behaviour.

References

- [1] Panagiotis T. Metaxas, Eni Mustafaraj Social Media and the Elections *Science*, 338:6106, pp.472-473, 2011.
- [2] C Lui, PT Metaxas, E Mustafaraj On the predictability of the US elections through search volume activity In: *Proceedings of the IADIS International e-Society* , 2011.
- [3] Maleki A. The latest estimation of the election turnout and votes distribution *BBC Persian*, http://www.bbc.co.uk/persian/iran/2013/06/130614_l45_ir92_polls_analysis.shtml , 2013.
- [4] Daniel Gayo-Avello A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data *Social Science Computer Review*, 2013.
- [5] Dutton WH., and Blank G. Next Generation Users: The Internet in Britain. Oxford Internet Survey 2011 Report. *Oxford Internet Institute, Oxford University*, 2011.
- [6] Choi H, Varian H. Predicting initial claims for unemployment benefits. *Available at* <http://research.google.com/archive/papers/initialclaimsUS.pdf>, 2009.
- [7] Choi H, Varian H. Predicting the present with Google Trends. *Available at* http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf, 2009.
- [8] Goel S., Hofman JM., Lahaie S., Pennock DM., Watts, DJ. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 107:41, pp17486, 2010.
- [9] Cook S, Conrad C, Fowlkes AL, Mohebbi MH Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE* 6(8): e23610, 2011.
- [10] Jungherr A. Tweets and Votes, a Special Relationship: The 2009 Federal Election in Germany. In: *Proceedings of the 2Nd Workshop on Politics, Elections and Data*, pp5-14, 2013.
- [11] O’Connor B, Balasubramanyan R, Routledge BR, et al. From tweets to polls: linking text sentiment to public opinion time series. In: *Proceedings of the fourth international AAAI conference on weblogs and social media*, Washington, DC, 23-26 May, 2010.
- [12] A. Ceron, L. Curini, and S. M. Iacus. 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* 2013.
- [13] Mislove A, Lehmann S, Ahn Y, Onnela J, Rosenquist J Understanding the demographics of Twitter users. In: *Fifth international AAAI conference on weblogs and social media* 2011.
- [14] Gayo-Avello D. I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper” – a balanced survey on election prediction using Twitter data. *preprint; arXiv:12046441* 2013.

- [15] Gayo-Avello D, Melaxas P, Mustafaraj E Limits of electoral predictions using Twitter. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. pp. 490-493. (2011)
- [16] Tumasjan A, Sprenger TO, Sander PG, Welppe IM. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. pp. 178-185 2010.
- [17] Mestyán, M.,Yasseri, T., Kertész J. Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE*, 8(8), 71226, 2013.
- [18] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, WWW '10, pp. 851–860, 2010.
- [19] Okazaki M, Matsuo Y. Semantic Twitter: Analyzing Tweets for real-time event notification. In: *Breslin J, Burg T, Kim HG, Raftery T, Schmidt JH, editors, Recent Trends and Developments in Social Software*, Springer, volume 6045 of Lecture Notes in Computer Science. pp. 63-74, 2011.
- [20] Lazer, David, et al. Computational Social Science. *Science*, 323, 5915:721-723, 2009.
- [21] Conte, R. et al. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214, 1:325-346, 2012.