(Wisdom of the Crowds)$^2$: 2010 UK Election Prediction with Social Media

FABIO FRANCH
Abstract

The vote share of the 2010 UK General Election is here forecasted by applying twice the concept behind Galton’s predictive ‘wisdom of the crowds’. Firstly, by aggregating at the media level (Facebook, Twitter, Twitter Sentiment, YouTube, Google) the political opinion of the audience. Secondly, by averaging at the media level each prediction. The ARIMA model performs predictions for the Labour party, Liberal-Democrats and Conservatives that are respectively only 0.48, 0.83 and 0.19 percentage points off the real vote share, thus well exceeding the predictive power of more traditional and expensive polls.
1. Introduction

Election forecasting has grown in importance within the Political Science field, creating a small, but fast expanding area of concentration for several political scientists and economists. American elections have been widely studied (Lewis-Beck & Rice, 1984; Campbell & Garand, 2000; Fishbein, & Hinkle, 1980; Abramowitz, 1988; Campbell, 2000; McDonald, 2010), but an increasing number of studies has focused on British elections (Nadeau, Lewis-Beck, Bélanger, 2009; Whiteley et al., 2010). This is still a small, but very active niche which has employed a variety of methods and techniques.

Whiteley (2005) reports that one branch of research in electoral forecasting for UK general elections has focused on the relationship between the economy and political support through popularity functions (Whiteley, 1986; Lewis-Beck, Nadeau and Bélanger, 2004; Sanders, 2005a/b), which concentrate on the effect of factors explaining support for incumbents. Other researchers have focused on the use of election data to estimate vote functions which could perform at least as well as traditional poll data (Lewis-Beck, Nadeau and Bélanger, 2004; Mughan, 1987). A smaller group of political scientists has focused on the use of election night data to predict election outcomes (Payne and Brown 1981) or forecasting seats in the House of Commons from vote shares (Kendall and Stuart, 1950; Qualter, 1968; Laakso, 1979).

__1__ STATA databases and a do file are available at [http://dvn.iq.harvard.edu/dvn/dv/jitp](http://dvn.iq.harvard.edu/dvn/dv/jitp) for replication purposes.

3 – (Wisdom of the Crowds)
The estimation of vote functions, started with Mughan (1987), has developed only in recent years, despite the relatively good results obtained by Mughan, whose economic model with lagged macroeconomic data as independent variables, performed better in the longer term than the incremental model, based exclusively on its own lagged value. The estimation of vote functions seemed to lose ground to the study of popularity functions until the beginning of the new millennium, when Lewis-Beck, Nadeau and Bélanger (2004), Norpoth (2004), Lebo and Norpoth (2006) published the results of their research. In the first case, using inflation and a measure of government approval the authors reach unsatisfactory results, despite a high Adjusted-R squared and a sound theory behind the variable selection. Their results were over 5 points off the actual vote share. Lebo and Norpoth (2006) used a second-order autoregressive model with just one independent variable: satisfaction towards the Prime Minister (PM). Noting that such measure is probably the only single variable which can capture with precision the short-term swing of electoral preferences, the authors build a model which appears to be significant in both theoretical terms and in prediction accuracy, with a 1.5 vote percentage mean error (Nadeau, Lewis-Beck and Bélanger, 2009). Lebo and Norpoth (2010) present a similar model for the 2010 General election, emphasizing the supremacy of a measure of PM approval over a measure of the Government approval for prediction purposes and stressing that “elections in Britain are certainly about choosing a Prime Minister, but they are thought to be first and foremost about choosing a government” (Lebo and Norpoth, p. 2).

Theoretical justification: Wisdom of the crowds and polls

4 – (Wisdom of the Crowds)²
An increasingly larger number of political scientists have been studying the predictive ability of political markets (Erikson and Wlezien, 2008; Berg et al., 2008; Gaissmaier and Marewski, 2011). Surowiecki’s (2004) ‘The wisdom of crowds’ is probably a milestone in this field in that it illustrates the predicting power of common people when their forecasts of uncertain quantities or future events are aggregated² (Gaissmaier and Marewski, 2011; Timmermann, 2006; Wolfers & Zitzewitz, 2004). According to his main idea, “boundedly rational individual” are capable of making, all together, a near-to-optimal decision, often outperforming every individual’s intelligence, meaning that the crowd, taken as an intelligent entity, is smarter than most of any human counterparts taken singularly. Google is a search engine that took advantage of this idea and was able to replace well-established competitors like Yahoo!, Altavista and Lycos. It did so simply by adopting a computational procedure which is able to let all Web pages on the Internet decide which page corresponds to the one that the user is looking for. There are limitations to this theory, and Surowiecki dedicates ample parts of his book thereon, explaining that a crowd, to be wise, needs to be diverse, independent, while its decisional procedure has to be decentralized³. The author proves his point describing how riots and stock market bubbles are evident cases where the lack of diversity, independence and decentralization leads a group to irrational decisions.

To access to a sample sufficiently large to contain diverse and independent pieces of information, one can resort to large, expensive and time-consuming political markets (Berg et al., 2008; Erikson & Wlezien, 2008). However, the Internet offers its users

² Surowiecki’s ‘wisdom of the crowds’ is usually explained with story of how Sir Francis Galton discovered that the aggregation (by average) of all the peasants’ guesses of an ox’s weight closely resembled the real figure (Galton, 1907).
³ See introduction.
ample opportunities to participate in a decisional process in a way that guarantees, at least to a certain degree, diversity, independence and decentralization, so that one can exploit the ‘wisdom of the crowds’ to perform accurate guesses on important political phenomena.

Social applications like Facebook, Twitter, YouTube and blogs offer access to a vast, diverse pool of potential interviewees. These applications, Facebook in particular, have become so popular that they are now integral part of the social life of large portions of our society’s wisdom (Madge et al., 2009). Moreover, these applications allow their users to freely express their opinions concerning any topic, including politics and preferred political party (Woodly, 2007). In particular, Facebook provides indirect ways to discover what the political preferences of the online community are by counting and comparing, for example, the number of Facebook friends a political leader has (Williams and Gulati, 2008; Upton, 2010). Other means of communications exist on this platform like chats, the creation of interest groups and gaming applications, all of which make this social network a very useful tool for electoral purposes as well.

Gregory Upton Jr (2010) finds that the number of Facebook friends is correlated with election outcomes. The more a person is attractive, the more Facebook friends he/she will have, the more likely it is that the individual successfully runs for an election. Williams and Gulati (2008) discover that Facebook support is a good measure of electoral success even after controlling for key political and demographic variables. The authors also find

---

4 Sjöberg (2009) has shown that the wisdom of the crowds can outperform the more traditional polls forecasts, simply by asking voters to rank running parties based on the number of votes the interviewed voter thinks these parties will receive. In line with Galton’s discovery, Sjöberg claims the average of all voters’ rankings can be then used to accurately forecast the winning party.


6 Their models explained 77-86% of the total variance.

6 – (Wisdom of the Crowds)^2
that this impact is more evident for those candidates who made a larger use of this social network for electoral purposes. No relationship was found between Facebook support and electoral success for those candidates who ignored social media as campaigning tools.

The predictive power of new social media is also studied by Tumasjan et al. (2010), who note that Twitter is a tool that is heavily used for political discussions. Their analyses and discussions emphasize the fact that the number of times a party is mentioned in Twitter messages is highly correlated with the success of that party. Moreover, they find that Tweet’s political sentiment well describes the current political landscape, meaning that Twitter messages can be successfully studied to predict electoral outcomes and have a valid snapshot of the political feeling of the public.

Shah (2010) developed a crawler tool which is able to collect a large amount of information from the YouTube community. A quick look at data collected for the video “Barack Obama: My Plans for 2008” revealed that on August 26, 2007 something special happened in the political arena. Further investigation revealed that Obama had just visited New Orleans, presenting a speech in which he advocated the reconstruction of the city and promoted a more expeditious response of the federal government to future disasters. This event, Obama’s walk in the city, triggered many discussions both in the news media and in the blogosphere (Shah, 2010, p. 10), which, in turn, was reflected by the change of behavior of the data collected concerning Obama’s promotional video.

Anvik and Gjelstad (2010) find that even Google has predictive powers. Using a Box Jenkins’ ARIMA model and data collected from Google Insights, which calculates how many times a specific term has been entered in the search engine in a specific geographical area, they note that Google searches have the ability to predict

---

7 See also Woodly (2007).

7 – (Wisdom of the Crowds)²
unemployment rates in the short term. The two authors conclude that despite the simplicity of the measures adopted, their models outperform the current leading unemployment indicator of reference (63).

Gloor et al. (2009) use social network analysis to mine the Web, blogs, and online forum with the intent of predicting long-term trends on various concepts, including on the popularity of politicians. Using a specific software, designed to find the most relevant people in these networks, to compute network centrality measures and to mine text-based data, the authors compare Web betweenness for the 2008 US presidential election Web buzz. They find that the temporal calculation of betweenness allows for the prediction of long-term trends of popularity (Gloor et al., p. 1) and show that “buzz on the Web mirrors the real world. Tracking concepts on the Web by differentiating between the Web at large, blogs, and online forums, and combining what people say with their social network position indeed permits to discover trends, frequently before the real world has become aware of them” (Gloor et al., p. 7).

Data

Given the evidence just discussed above that tapping the ‘wisdom of the crowds’ with the new media may allow for the construction of an accurate electoral prediction model, several independent variables were employed to measure the approval of the future Prime Minister. In line with Lebo and Norpoth (2006; 2010) work, measures of PM approval were adopted, approval which, alone, is the variable can be used to accurately predict an electoral outcome.

---

8 The term ‘betweenness’ refers to the property of being between different pairs or sets of other entities in a particular network.
Measures of PM popularity were derived from the usage of Facebook, Twitter, Google and YouTube, in line with the pieces of literature described above. The goal of this paper is that of exploiting to its maximum the predictive power of the ‘wisdom of the crowds’. That is achieved not only by resorting to tapping it by finding indicators of expression of political belief in the new media. The power of the ‘wisdom of the crowds’ is also maximized by analyzing several of these media (Facebook, Twitter, Google, YouTube), with the intent of avoiding potential biases stemming from an excessive reliance on just one media. The positive properties of the ‘wisdom of the crowds’ are in this study used twice, at two different levels. One time to aggregate the opinions of hundreds or thousands of individuals using a specific media, and another time to aggregate these political believes at the media level (by averaging the significant predictors).

The data have been manually collected from Facebook, Twitter Sentiment, the Twitter-based Topsy, Google, and YouTube. With regards of Facebook, the largest group/fan page (in terms of number of people that like it) dedicated to support each of the three main political leaders is selected. In addition, one Facebook group which expresses its opposition to each leader as well is chosen. Twitter was used in order to

---

9 This is achieved by averaging all media predictions.


have a measure of how popular each of the political leaders was during a day. The Twitter-based application Topsy\textsuperscript{12} was employed for this purpose.

Google blogs provided a similar measure, by counting the number of times each political leader was named in blogs in the last 24 hours.

The popularity of each candidate was also measured through the number of visualizations of the official promotional video posted in YouTube. Where several options were present, the most popular (viewed the most times), was selected as representative.

Drawing on the literature presented above, it is so far assumed that it is enough to count how many times a candidate was named in a discussion or was presented in a YouTube video to accurately trace his popularity. In order to measure how well or badly each candidate is perceived by the public an additional measure was introduced using the application Twitter Sentiment\textsuperscript{13} and collecting data relative to the main political leaders: Gordon Brown, David Cameron and Nick Clegg.

Lastly, the predictions of a website dedicated to the 2010 General Election\textsuperscript{14} were included as an additional way to tap the predictive power of the crowds.

These variables have been processed in order to take into account other electoral and voting aspects. For example, given the larger popularity on the Internet enjoyed by Gordon Brown, the percentage increase (and not just the actual count) was also added to the list of independent variable to be regressed, each at a time, on the YouGov data.

\textsuperscript{12} \url{http://topsy.com/}
\textsuperscript{13} See \url{http://twittersentiment.appspot.com/}.
\textsuperscript{14} See \url{http://www.general-election-2010.co.uk}.
Moreover, as noted by Lebo and Norpoth (2006; 2010), who assert that British elections are largely about choosing a Prime Minister over another, a weighted measure for all the variables just described was also included in the list of independent variables, for a grand total of 25 independent variables\textsuperscript{15}. The Appendix contains a detailed explanation of all these variables, which is here omitted in order not to distract the reader. It is important to note that to avoid issues of multicollinearity, each statistical model was run with one dependent variable (the YouGov popularity function for each candidate) and one of the 25 independent variables described in depth in the Appendix and Tables IV, V and VI. The equation being estimated in every model can be represented as follows:

\[ Y_t = IV_t + e_t, \]

Where \( Y_t \) is the YouGov measurement at time \( t \) of the vote share for the Labour Party the Conservatives, or the Liberal Democrats, \( IV_t \) is one the 25 independent variables presented in the Appendix (the significant ones have been reported in Table I, II, and III) and \( e_t \) is the error.

The logic behind regressing each created variable on the YouGov is based on the belief that every variable represents an alternative expression of political preference, similar to voting. Regressing more than one at a time would certainly mean creating multicollinearity, which in turn means having higher standard errors and thus a higher

\textsuperscript{15} Poll results from the GE2010 website, Facebook friends, Facebook friends percentage increase, Facebook friends weighted percentage increase, number of members of the largest opposing Facebook group, number, number of members of the largest opposing Facebook group percent increase, weighted number of members of the largest opposing Facebook group percent increase, times each candidate was Twitted, its percentage increase, Twitter Sentiment (positive) count and its percentage increase, Twitter Sentiment (negative) count, its percentage increase and its weighted percentage increase, the difference between the positive and negative count, its percentage increase and its weighted percentage increase, percentage difference between positive and negative over positive, weighted percentage difference between positive and negative over positive. Other variables were derived from the YouTube view of the official promotional video, its percentage and its weighted percent increase, the count of blogs mentioning each candidate’s name as measured by Google, percentage increase and weighted percentage increase. See the Appendix for a more detailed description and definition of these variables.
probability of accepting a false hypothesis that an independent variable can help predict the dependent variable.

The data collected range from April 7, 2010 to May 6, 2010, the election day. The estimation of the model is exclusively based only on the period prior to the election day. The election day is here used to make out-of-sample forecasts.

The dependent variable, YouGov polls data\textsuperscript{16}, has been chosen exclusively for its daily frequency\textsuperscript{17}. Given the short frame during which the data was available for collection (the election day was confirmed on April 6, 2010, just one month before the election), YouGov poll data appeared to be the optimal solution. The data were collected daily at around 11PM Eastern Time Zone.

**Methodology**

In this study, I measure the predicting power of social media regressing one variable at a time using an Auto Regressive Integrated Moving Average (ARIMA) model, also known as Box-Jenkins model\textsuperscript{18}. These models are based on the observation that with time-series data the dependent variable $Y_t$ can be explained not only by correctly specified independent variables, but also by the previous values (in temporal terms) of the dependent variable $Y_{t-1}$ and values of the independent variables that do not refer to time $t$. In other words, ARIMA models can be used to properly deal with the problem that the residuals of the errors terms are correlated with previous time lags and can be used to study phenomena that have evolved over time. Box and Jenkins should not be credited for the creation of all the elements of the ARIMA models. However, they should be credited for their comprehensive formulation of the ARIMA model (McDowall et al., 1980, p. 14).

\textsuperscript{16} Data collected from http://ukpollingreport.co.uk/blog/.

\textsuperscript{17} Angus Reid would have probably been a better dependent variable, given its higher accuracy.

\textsuperscript{18} ARIMA models are based on the observation that with time-series data the dependent variable $Y_t$ can be explained not only by correctly specified independent variables, but also by the previous values (in temporal terms) of the dependent variable $Y_{t-1}$ and values of the independent variables that do not refer to the time $t$. In other words, ARIMA models can be used to properly deal with the problem that the residuals of the errors terms are correlated with previous time lags and can be used to study phenomena that have evolved over time. Box and Jenkins should not be credited for the creation of all the elements of the ARIMA models. However, they should be credited for their comprehensive formulation of the ARIMA model (McDowall et al., 1980, p. 14).
The predictions of all significant variables were then averaged, thus performing an analysis that closely resembles Galton’s.

In order to correctly specify the model, the data set was plotted in order to confirm the stationarity of the data. At this point an upward or downward trend was detected which required differentiation for all the models predicting the vote percentage of the Liberal Democrats, the Conservatives and the Labour party (Figure 1).

### FIGURE 1

A first-order autoregressive process was indentified in all the models through the observation of the AutoCorrelation (AC) and Partial AutoCorrelation (PAC) functions and the residuals\(^{19}\), thus deviating from the models used by Lebo and Norpoth (2006). The Portmanteau test for white noise consistently confirmed the absence of residual processes in all models, whose goodness-of-fit was further supported by a lower value returned by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) tests\(^{20}\).

### Results

A quick look at the data reveals that the data collected closely follow the trend showed by YouGov polls data. An example is presented in Figure 2, where the online poll data collected from the 2010 General Election website are juxtaposed to the official YouGov poll data. Both refer to the statistics relative to the Labour Party.

\(^{19}\) Autoregressive processes have decaying ACF graphs and spiking PACF graphs. See Cryer and Chan (2008) to know more modeling a time series.

\(^{20}\) These tests indicate the relative goodness of fit of a statistical model based on their informational loss and are thus useful for model selection.
FIGURE 2

Not only social media can capture the general trend in the public opinion, as demonstrated in Figure 2. They have also the power to capture in real time swing in the public opinion after specific political events.

FIGURE 3

In Figure 3 one can observe that in correspondence with the Gordon Brown’s ‘sinking interview’ on April 18, 201021 the percent increase in the largest Facebook group opposing Gordon Brown experienced a sudden surge, suggesting a negative reaction of the public to Gordon Brown’s statements. A similar event was recorded on April 28, 2010, when Gordon Brown was caught calling a Labour supporter ‘bigoted woman’22. The same data show that on the occasion of the April 22 Leaders’ debate, the public reacted joining Facebook groups opposing the Labour leader, even though Gordon Brown’s reputation came out of the debate slightly improved.

Forecasts and estimations based on each independent variable confirm the power of social media at tracing political opinion swings and predicting electoral outcomes. Only few of the all the independent variables selected are found to be significant (p<0.05). Nonetheless, their average prediction is surprisingly very close to the real electoral outcome. Please note that in Tables I, II, and III significance levels and each regression’s Chi-Square is presented. Chi-Square values close to zero leave little doubt that such results are due to randomness stemming from the small sample size (1 independent variable and N=30).

TABLE I

Table I shows that when predicting the Conservatives vote percentage, the Facebook Fans Percentage Increase (FFPI), the number of Facebook members of opposing group Percentage Increase (FDPI), Twitter Sentiment Positive (TSP), Twitter Sentiment Positive Percent Increase (TPPINC) and Twitter Sentiment Negative Percent Increase (TNPINC) are significantly related\textsuperscript{23} to the YouGov variable, showing that the positive/negative context of reference to the leader are important factors when it comes to election prediction outcomes. For the last two models, the Chi-Square value surpasses the traditionally accepted value of 0.05, meaning that there is a slightly higher probability that such predictions are due to randomness, rather than model accuracy. However, these models’ predictions seem to be in line with the others, meaning that such problematic may be overstated.

The average prediction based on all these variables closely matches the real electoral outcome (0.4849 percentage point error) supporting the power of the second application of the ‘wisdom of the crowds’ at the media level. These results, however, are not robust for the forecasts based on the regression up to the second-to-last day and third-to-last day before the election day, in that they return respectively an average prediction equal to 34.298788 and 34.265892. This suggests that in the case of the Conservative party the last 2 and 3 days were rather eventful to the point where the statistical model loses some of its ability to predict the correct electoral outcome. A regression based on data which do not include these days seems to lose some of its predicting power.

**TABLE II**

Similar variables (Facebook Fans Percentage Increase – FFPInc, Facebook Fan Weighted Percentage Increase – FFWPI, Facebook members of opposing group Percentage

\textsuperscript{23}95% confidence interval.

15 – (Wisdom of the Crowds)\textsuperscript{2}
Increase – FDPI) but a larger forecast error (+0.8273 percentage points) characterizes the Liberal-Democratic vote share prediction (see Table II). It is here that the model presented in this paper shows its real power, with an improvement over the YouGov forecast much bigger and significant than for the Conservatives and the Labour Party (see below). Once again, these results based on the regression up to the second-to-last day and third-to-last day before the election day return respectively an average prediction equal to 23.18733 and 26.28944, suggesting that in the case of the Liberal-Democrats something had happened that affected the final vote especially on the third-to-last day before the election which has caused the model to predict a final vote share which is rather far away from the real figure (+3.28944).

**TABLE III**

Only one variable out of 25 (Facebook Fan Weighted Percentage Increase - FFWPI) is found to be a significant predictor of the Labour Party vote share in the 2010 UK General Election (Table III). Nonetheless, this variable predicts almost perfectly the real outcome, with only a 0.19338 percentage point error. These forecasts are rather consistent with the forecasts based on the regression up to the second-to-last day and third-to-last day before the election day, which return, respectively, an average prediction equal to 29.75584 and 28.46665, thus showing a small decrease in accuracy for both models.

**Conclusions**
The intent of this study is not that of explaining political voting, rather it is that of using social media to predict election outcomes\textsuperscript{24}. In line with Shah (2010) and Gloor et al. (2009) findings, social media are here found to be able to effectively aggregate public opinions regarding political matters and to be reliable measures of public opinion swings.

The accuracy of the model developed here has exceeded any expectation and the accuracy of other models found in the literature, not to mention all opinion polls (see Table IV).

**TABLE IV**

Despite the fact that polls are well known for being based on unrepresentative samples (which can bias the estimation of the regression coefficients and thus the predicted vote share), the ARIMA model seems to bypass such flaw and leads to forecasts that surpass in accuracy those of the more traditional and expensive polls. These accurate forecasts cannot be explained even in the face of potential sample bias issues (Wei and Hindman, 2011), in that the Internet tends to be more heavily used by highly educated, male individuals\textsuperscript{25}. On this regards, Bakker and de Vreese (2011) have suggested that even though Internet use is more widespread in younger generations, their lower tendency to participate in political matters might contribute to make political participation more evenly distributed among age cohorts.

The consequences of this study are important. As Xin et al. (2010) note, models based on the social media are strong competitors of empirical surveys, which are affected by sampling bias, method bias and higher costs (Xin et al., p. 4). The model presented not

\textsuperscript{24} The reader needs to note that these forecasts are based on an independent variable, meaning that it is necessary to have a measure of such variable to have a reliable estimate of the current vote share of each party. No forecast is here discussed which is solely based on the dependent variable.

\textsuperscript{25} Age, Race and Income are found non-significant. Please note that these results apply to the US.
only outperforms all opinion polls, but it is inexpensive and can reach a wide audience essentially in real time and in a less complex fashion. The polling industry should adjust to take into account the knowledge of modern ‘wisdom of the crowds’.

This study could benefit from a larger sample, limited, however, by the short period between the announcement of the electoral confirmation and the election day itself. Nonetheless, the Chi-Square measure reported for each model shows that sample size is, in most cases, unlikely to be the reason behind the accuracy shown. Even for the two models whose Chi-Square value is higher than normal, consistency with the other forecasts seems to support the hypothesis that the limited number of observations plays little role in explaining the goodness of this approach. Future research will entail the automatic collection of similar data via Web crawlers in the attempt to predict the winner of the 2012 US election using a similar theoretical and empirical framework. Such research, if successful, will add to the robustness of the findings, thus further reducing the likelihood that randomness is behind the accuracy of the results shown here.

---

26 Done especially for comparative purposes.
References


20 – (Wisdom of the Crowds)$^2$


Madge, Clare, Julia Meek, Jane Wellens, Tristram Hooley. (2009). Facebook, social integration and informal learning at university: It is more for socialising and talking to friends about work than for actually doing work. *Learning, Media and Technology*, 34 (2), 141-155. doi:10.1080/17439880902923606.


Appendix

The variables considered and included in this study and built for all three Parties are defined as follows:

1. **GE2010 website vote share (GE2010)** (http://www.general-election-2010.co.uk/general-election-2010-economy-polls.html): actual vote over the total votes expressed up to that point.

2. **Facebook Friends (FF):** for Labour Party, the number of fans was taken from http://www.facebook.com/search/?flt=1&q=gordon+brown&o=69&sid=571193786.1775585227..1&s=10#!/pages/Gordon-Brown/67132943785?ref=ts simply by looking at the raw number shown in the Facebook page. For the Conservative Party, a similar piece of information was taken from http://www.facebook.com/DavidCameron?ref=ts#!/DavidCameron?v=wall&ref=tss, while for the Liberal-Democratic Party the Facebook group found at http://www.facebook.com/nickclegg?ref=search&sid=571193786.2593503385..1 was used for the purpose.

3. **Facebook Friends Percentage Increase (FFPInc):** in order to compensate for leader’s popularity issues (the assumption is that Gordon Brown has enjoyed a larger popularity up to the General Election), the percentage increase was used,
where $\text{FFPI}_{\text{c}} = (\text{FF}_{t} - \text{FF}_{t-1}) / \text{FF}_{t-1}$. A similar logic is adopted for all measures presented below and expressed as Percentage Increase.

4. **Facebook Friends Weighted Percentage Increase (FFWPI)**: measure introduced due to the belief that in an election the candidate who performs better relative to the others has a better chance to win. A similar logic is adopted for all weighted measures presented below. Defined as follows:

$$\text{FFWPI}_L = \frac{\text{FFPI}_{\text{c}}}{(\text{FFPI}_{\text{c}} + \text{FFPI}_{\text{c}} + \text{FFPI}_{\text{LD}})},$$

$$\text{FFWPI}_C = \frac{\text{FFPI}_{\text{c}}}{(\text{FFPI}_{\text{c}} + \text{FFPI}_{\text{c}} + \text{FFPI}_{\text{LD}})},$$

$$\text{FFWPI}_{\text{LD}} = \frac{\text{FFPI}_{\text{c}}}{(\text{FFPI}_{\text{c}} + \text{FFPI}_{\text{c}} + \text{FFPI}_{\text{LD}})},$$

respectively for the Labour (L), Conservative (C), and Liberal Democratic Party (LD). Computed for every day of the campaign.


26 – (Wisdom of the Crowds)²
6. **Members of the largest opposing Facebook group percent increase (FDPI):**

   measure calculates as for FFPIInc. Defined as:

   $$\text{FDPI} = (\text{FLDG}_t - \text{FLDG}_{t-1}) / \text{FLDG}_{t-1}.$$

7. **Weighted number of members of the largest opposing Facebook group percent increase (FDWPInc):** measure calculated following the same logic behind FFWPI. Defined as:

   $$\text{FDWPInc}_L = \text{FDPI}_L / (\text{FDPI}_L + \text{FDPI}_C + \text{FDPI}_{LD}),$$
   $$\text{FDWPInc}_C = \text{FDPI}_C / (\text{FDPI}_L + \text{FDPI}_C + \text{FDPI}_{LD}),$$
   $$\text{FDWPInc}_{LD} = \text{FDPI}_{LD} / (\text{FDPI}_L + \text{FDPI}_C + \text{FDPI}_{LD}),$$ respectively for the Labour, Conservative, and Liberal Democratic Party. Computed for every day of the campaign.


9. **Twitter Percentage Increase (TPInc):** measure defined as follows:

   $$\text{TPInc} = (T_t - T_{t-1}) / T_{t-1}$$


11. **Twitter Sentiment - Positive percentage increase (TPPInc):** defined as

    $$\text{TPPInc} = (\text{TSP}_t - \text{TSP}_{t-1}) / \text{TSP}_{t-1}.$$

13. Twitter Sentiment – Negative Percentage Increase (TNPInc):

\[ \text{TNPInc} = (\text{TSN}_t - \text{TSN}_{t-1}) / \text{TSN}_{t-1}. \]

14. Twitter Sentiment – Negative weighted percentage increase (WTNPInc):

calculated as follows for every day of the campaign:

\[ \text{WTNPInc}_L = \text{TNPInc}_L / (\text{TNPInc}_L + \text{TNPInc}_C + \text{TNPInc}_{LD}), \]

\[ \text{WTNPInc}_C = \text{TNPInc}_C / (\text{TNPInc}_L + \text{TNPInc}_C + \text{TNPInc}_{LD}), \]

\[ \text{WTNPInc}_{LD} = \text{TNPInc}_{LD} / (\text{TNPInc}_L + \text{TNPInc}_C + \text{TNPInc}_{LD}), \]

respectively for the Labour, Conservative, and Liberal Democratic Party.

15. Difference between the Twitter Positive and Negative count (TDPN): computed simply by taking the difference between TSP and TSN.

16. Difference between the Twitter Positive and Negative percentage increase (TDPNPInc):

\[ \text{TDPNPInc} = (\text{TDPN}_t - \text{TDPN}_{t-1}) / \text{TDPN}_{t-1}. \]

17. Difference between the Twitter Positive and Negative weighted percentage increase (TWPDPNPInc):

calculated as follows for every day of the campaign:

\[ \text{TWPDPNPInc}_L = \text{TDPNPInc}_L / (\text{TDPNPInc}_L + \text{TDPNPInc}_C + \text{TDPNPInc}_{LD}), \]

\[ \text{TWPDPNPInc}_C = \text{TDPNPInc}_C / (\text{TDPNPInc}_L + \text{TDPNPInc}_C + \text{TDPNPInc}_{LD}), \]

\[ \text{TWPDPNPInc}_{LD} = \text{TDPNPInc}_{LD} / (\text{TDPNPInc}_L + \text{TDPNPInc}_C + \text{TDPNPInc}_{LD}), \]

respectively for the Labour, Conservative, and Liberal Democratic Party.
18. **Percentage Difference between Positive and Negative over Positive** (TPDPS):
   defined as: \( TPDPS = \frac{(TSP - TSN)}{TSP} \).

19. **Weighted Percentage Difference between Positive and Negative over Positive** (TWPDPN): calculated as follows for every day of the campaign:
   
   \[
   TWPDPN_L = \frac{TPDPS_L}{(TPDPS_L + TPDPS_C + TPDPS_{LD})},
   \]
   
   \[
   TWPDPN_C = \frac{TPDPS_C}{(TPDPS_L + TPDPS_C + TPDPS_{LD})},
   \]
   
   \[
   TWPDPN_{LD} = \frac{TDPNI_{LD}}{(TPDPS_L + TPDPS_C + TPDPS_{LD})},
   \]
   respectively for the Labour, Conservative, and Liberal Democratic Party.

20. **YouTube views of the official promotional video** (YouTube): number of times the official video was viewed.

21. **YouTube views percentage increase** (YPInc):
   \[
   YPInc = \frac{(YouTube_t - YouTube_{t-1})}{YouTube_{t-1}}.
   \]

22. **YouTube views weighted percent increase** (YWPinc): calculated as follows for every day of the campaign:
   
   \[
   YWPinc_L = \frac{YPInc_L}{(YPInc_L + YPInc_C + YPInc_{LD})},
   \]
   
   \[
   YWPinc_C = \frac{YPInc_C}{(YPInc_L + YPInc_C + YPInc_{LD})},
   \]
   
   \[
   YWPinc_{LD} = \frac{YPI_{LD}}{(YPInc_L + YPInc_C + YPInc_{LD})},
   \]
   respectively for the Labour, Conservative, and Liberal Democratic Party.

23. **Number of Blogs mentioning each candidate’s name as measured by Google** (Gblogs): number returned using Google, selecting “Blogs” from the left menu and using “‘CANDIDATE NAME’ winner’ as key words, where CANDIDATE
NAME is the name of each running candidates. The number used in this study is the one returned right under the search engine box.

24. *Number of Blogs percentage increase* (**GPInc**):

\[ \text{GPInc} = \frac{\text{Gblogs}_t - \text{Gblogs}_{t-1}}{\text{Gblogs}_{t-1}}. \]

25. *Number of Blogs weighted percentage increase* (**GWPinc**): calculated as follows for every day of the campaign:

\[ \text{GWPinc}_L = \frac{\text{GPInc}_L}{(\text{GPInc}_L + \text{GPInc}_C + \text{GPInc}_{LD})}, \]

\[ \text{GWPinc}_C = \frac{\text{GPInc}_C}{(\text{GPInc}_L + \text{GPInc}_C + \text{GPInc}_{LD})}, \]

\[ \text{GWPinc}_{LD} = \frac{\text{GPInc}_{LD}}{(\text{GPInc}_L + \text{GPInc}_C + \text{GPInc}_{LD})}, \]

respectively for the Labour, Conservative, and Liberal Democratic Party.

Below, in Tables V, VI and VII, the summary statistics are reported relative to the dependent and independent variables.

| Table V: Summary statistics of dependent and independent variable (Labour Party) |
|---------------------------------|------|---------|-----------|---------|------|
| Variable                      | Observations | Mean       | Standard Deviation | Min       | Max      |
| YOUGOV                         | 30       | 28.96667  | 1.93842            | 26        | 33       |
| GE2010                         | 30       | 0.224     | 0.01765            | 0.205903  | 0.250518 |
| FF                             | 30       | 7336.767  | 1278.85            | 5241      | 9493     |
| FFPINC                         | 29       | 0.020758  | 0.011591           | 0.004254  | 0.062674 |
| FFWPI                          | 29       | 0.214984  | 0.096101           | 0.039864  | 0.381226 |
| FLDG                           | 30       | 33677.07  | 2838.19            | 29631     | 38661    |
| FDPI                           | 29       | 0.009224  | 0.004443           | 0.004015  | 0.018271 |

30 – (Wisdom of the Crowds)$^2$
<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDWPINC</td>
<td>29</td>
<td>0.078952</td>
<td>0.060923</td>
<td>0.004091</td>
<td>0.215622</td>
</tr>
<tr>
<td>T</td>
<td>30</td>
<td>1004.767</td>
<td>901.653</td>
<td>332</td>
<td>4386</td>
</tr>
<tr>
<td>TPINC</td>
<td>29</td>
<td>0.393259</td>
<td>1.98865</td>
<td>-0.61013</td>
<td>10.36269</td>
</tr>
<tr>
<td>TSP</td>
<td>30</td>
<td>4476.067</td>
<td>1629.944</td>
<td>2403</td>
<td>7688</td>
</tr>
<tr>
<td>TPPINC</td>
<td>29</td>
<td>0.042353</td>
<td>0.058271</td>
<td>-0.00031</td>
<td>0.251801</td>
</tr>
<tr>
<td>TSN</td>
<td>30</td>
<td>6860.2</td>
<td>2735.038</td>
<td>3534</td>
<td>12121</td>
</tr>
<tr>
<td>TNPINC</td>
<td>29</td>
<td>0.045356</td>
<td>0.068561</td>
<td>0.005583</td>
<td>0.299354</td>
</tr>
<tr>
<td>WTNPINC</td>
<td>29</td>
<td>0.070918</td>
<td>0.271462</td>
<td>-0.21303</td>
<td>1</td>
</tr>
<tr>
<td>TDPN</td>
<td>30</td>
<td>-2384.13</td>
<td>1113.041</td>
<td>-4433</td>
<td>-1131</td>
</tr>
<tr>
<td>TDPNPINC</td>
<td>29</td>
<td>-0.05175</td>
<td>0.095831</td>
<td>-0.47909</td>
<td>-0.00751</td>
</tr>
<tr>
<td>TWPDPNPINC</td>
<td>20</td>
<td>-0.00882</td>
<td>0.052913</td>
<td>-0.15426</td>
<td>0.129683</td>
</tr>
<tr>
<td>TPDPS</td>
<td>30</td>
<td>-0.51595</td>
<td>0.055067</td>
<td>-0.61605</td>
<td>-0.47025</td>
</tr>
<tr>
<td>TWPDPN</td>
<td>29</td>
<td>-1.36306</td>
<td>5.594163</td>
<td>-29.0687</td>
<td>1.213529</td>
</tr>
<tr>
<td>YOUTUBE</td>
<td>30</td>
<td>402586.2</td>
<td>14501.4</td>
<td>385836</td>
<td>430200</td>
</tr>
<tr>
<td>YPINC</td>
<td>29</td>
<td>0.003763</td>
<td>0.002585</td>
<td>0.000536</td>
<td>0.009873</td>
</tr>
<tr>
<td>YWPINC</td>
<td>29</td>
<td>0.059936</td>
<td>0.049339</td>
<td>0.001183</td>
<td>0.160893</td>
</tr>
<tr>
<td>GBLOGS</td>
<td>30</td>
<td>640.4</td>
<td>329.9016</td>
<td>254</td>
<td>1430</td>
</tr>
<tr>
<td>GPINC</td>
<td>29</td>
<td>0.181585</td>
<td>0.581662</td>
<td>-0.68952</td>
<td>2.153543</td>
</tr>
<tr>
<td>GWPINC</td>
<td>29</td>
<td>0.128755</td>
<td>0.769764</td>
<td>-3.32561</td>
<td>1.198734</td>
</tr>
</tbody>
</table>

**Table VI: Summary statistics of dependent and independent variable (Conservatives)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOUGOV</td>
<td>30</td>
<td>35.33333</td>
<td>2.682425</td>
<td>31</td>
<td>41</td>
</tr>
<tr>
<td>GE2010</td>
<td>30</td>
<td>0.273693</td>
<td>0.022429</td>
<td>0.253746</td>
<td>0.31003</td>
</tr>
<tr>
<td>FF</td>
<td>30</td>
<td>28371.6</td>
<td>5552.957</td>
<td>20300</td>
<td>39296</td>
</tr>
<tr>
<td>FFPINC</td>
<td>29</td>
<td>0.023095</td>
<td>0.011078</td>
<td>0.0066</td>
<td>0.060478</td>
</tr>
<tr>
<td>FFWPI</td>
<td>29</td>
<td>0.243846</td>
<td>0.101952</td>
<td>0.044367</td>
<td>0.481894</td>
</tr>
</tbody>
</table>

31 – (Wisdom of the Crowds)²
Table VI: Summary statistics of dependent and independent variable (Liberal Democrats)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOUGOV</td>
<td>30</td>
<td>26.3</td>
<td>5.018243</td>
<td>18</td>
<td>34</td>
</tr>
<tr>
<td>GE2010</td>
<td>30</td>
<td>0.24397</td>
<td>0.073756</td>
<td>0.062531</td>
<td>0.307629</td>
</tr>
<tr>
<td>FF</td>
<td>30</td>
<td>32894.13</td>
<td>20567.22</td>
<td>4985</td>
<td>61246</td>
</tr>
</tbody>
</table>

Table VII: Summary statistics of dependent and independent variable (Liberal Democrats)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOUGOV</td>
<td>30</td>
<td>26.3</td>
<td>5.018243</td>
<td>18</td>
<td>34</td>
</tr>
<tr>
<td>GE2010</td>
<td>30</td>
<td>0.24397</td>
<td>0.073756</td>
<td>0.062531</td>
<td>0.307629</td>
</tr>
<tr>
<td>FF</td>
<td>30</td>
<td>32894.13</td>
<td>20567.22</td>
<td>4985</td>
<td>61246</td>
</tr>
</tbody>
</table>

32 – (Wisdom of the Crowds)²
<table>
<thead>
<tr>
<th>Variable</th>
<th>Month</th>
<th>Mean 1</th>
<th>Mean 2</th>
<th>Mean 3</th>
<th>Mean 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFPINC</td>
<td>29</td>
<td>0.101696</td>
<td>0.185434</td>
<td>0.010144</td>
<td>0.869725</td>
</tr>
<tr>
<td>FFWPI</td>
<td>29</td>
<td>0.54117</td>
<td>0.173834</td>
<td>0.288812</td>
<td>0.897396</td>
</tr>
<tr>
<td>FLDG</td>
<td>30</td>
<td>1943.033</td>
<td>1743.298</td>
<td>46</td>
<td>4761</td>
</tr>
<tr>
<td>FDPI</td>
<td>29</td>
<td>0.190591</td>
<td>0.225389</td>
<td>0.010905</td>
<td>0.978723</td>
</tr>
<tr>
<td>FDWPINC</td>
<td>29</td>
<td>0.737351</td>
<td>0.173516</td>
<td>0.329322</td>
<td>0.955647</td>
</tr>
<tr>
<td>T</td>
<td>30</td>
<td>784.9667</td>
<td>760.5866</td>
<td>119</td>
<td>3971</td>
</tr>
<tr>
<td>TPINC</td>
<td>29</td>
<td>0.306692</td>
<td>1.149979</td>
<td>-0.69151</td>
<td>5.339286</td>
</tr>
<tr>
<td>TSP</td>
<td>30</td>
<td>6.833333</td>
<td>2.972092</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>TPPINC</td>
<td>29</td>
<td>0.386324</td>
<td>1.410678</td>
<td>-0.83333</td>
<td>7</td>
</tr>
<tr>
<td>TSN</td>
<td>30</td>
<td>6.966667</td>
<td>3.112636</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>TNPINC</td>
<td>29</td>
<td>0.271945</td>
<td>0.916533</td>
<td>-0.90909</td>
<td>3</td>
</tr>
<tr>
<td>WTPINC</td>
<td>29</td>
<td>1.04148</td>
<td>2.60336</td>
<td>-2.74343</td>
<td>11.72866</td>
</tr>
<tr>
<td>TDPN</td>
<td>30</td>
<td>-0.13333</td>
<td>4.032055</td>
<td>-9</td>
<td>8</td>
</tr>
<tr>
<td>TDPNPINC</td>
<td>25</td>
<td>0.300294</td>
<td>2.184675</td>
<td>-4.5</td>
<td>5</td>
</tr>
<tr>
<td>TWPDPNPINC</td>
<td>20</td>
<td>0.077587</td>
<td>2.230211</td>
<td>-8.07982</td>
<td>2.042716</td>
</tr>
<tr>
<td>TPDPS</td>
<td>30</td>
<td>-0.29031</td>
<td>1.022192</td>
<td>-4</td>
<td>0.833333</td>
</tr>
<tr>
<td>TWPDPN</td>
<td>29</td>
<td>1.560141</td>
<td>5.573987</td>
<td>-0.82048</td>
<td>30.0687</td>
</tr>
<tr>
<td>YOUTUBE</td>
<td>30</td>
<td>1590.133</td>
<td>593.436</td>
<td>222</td>
<td>2395</td>
</tr>
<tr>
<td>YPINC</td>
<td>29</td>
<td>0.103211</td>
<td>0.25908</td>
<td>0.005065</td>
<td>1.418919</td>
</tr>
<tr>
<td>YWPINC</td>
<td>29</td>
<td>0.525588</td>
<td>0.180794</td>
<td>0.169461</td>
<td>0.874972</td>
</tr>
<tr>
<td>GBLOGS</td>
<td>30</td>
<td>560.9667</td>
<td>402.0981</td>
<td>8</td>
<td>1480</td>
</tr>
<tr>
<td>GPINC</td>
<td>29</td>
<td>0.506661</td>
<td>1.266699</td>
<td>-0.95699</td>
<td>5.479592</td>
</tr>
<tr>
<td>GWPINC</td>
<td>29</td>
<td>0.503047</td>
<td>0.589244</td>
<td>-1.24352</td>
<td>2.723623</td>
</tr>
</tbody>
</table>
Figure 1: Trends in the dependent variables for all 3 main Parties

- Labour
- Conservative
- Liberal Democrat
Figure 2: A close relationship between YouGov polls and the GE2010 online survey

(Labour party)
Figure 3: Gordon Brown's popularity against three different events (the 'sinking interview', a public debate, "bigoted woman") and against Facebook disapproval
Table I: Significant independent variables and forecast on election day (David Cameron)

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>FFPI</th>
<th>FDPI</th>
<th>TSP</th>
<th>TPPINC</th>
<th>TNPINC</th>
<th>Real outcome</th>
<th>Model Chi-square</th>
<th>P-value</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0102</td>
<td>0.0494</td>
<td>0.0165</td>
<td>0.0802</td>
<td>0.0747</td>
<td>35.62</td>
<td>0.003</td>
<td>0.014</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>32.6898</td>
<td>34.987</td>
<td>37.6500</td>
<td>36.89814</td>
<td>35.8501</td>
<td>35.62</td>
<td>0.003</td>
<td>0.014</td>
<td>0.025</td>
</tr>
</tbody>
</table>
Table II: Significant Independent variables and forecast on election day (Nick Clegg)

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>FFPInc</th>
<th>FFWPI</th>
<th>FDPI</th>
<th>Real</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Chi-Square</td>
<td>&lt;0.000</td>
<td>0.0002</td>
<td>&lt;0.000</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>&lt;0.000</td>
<td>&lt;0.000</td>
<td>&lt;0.000</td>
<td></td>
</tr>
<tr>
<td>Prediction</td>
<td>24.2233</td>
<td>22.9632</td>
<td>24.29537</td>
<td>23.82729</td>
</tr>
</tbody>
</table>

Average outcome
Table III: Significant independent variables and their forecast on election day

(Gordon Brown)

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>FFWPI</th>
<th>Real</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Chi-Square</td>
<td>0.0001</td>
<td>Average</td>
</tr>
<tr>
<td>p-value</td>
<td>0.031</td>
<td>outcome</td>
</tr>
<tr>
<td>Prediction</td>
<td>29.19338</td>
<td>29</td>
</tr>
</tbody>
</table>
Table IV: Official predictions of the main polling companies for Conservatives, Liberal Democrats, and Labour

<table>
<thead>
<tr>
<th>YouGov</th>
<th>ComRes</th>
<th>ICM</th>
<th>Angus Reid</th>
<th>Populus</th>
<th>MORI</th>
<th>Average</th>
</tr>
</thead>
</table>