Xie et al. *EPJ Data Science* (2018) 7:32
https://doi.org/10.1140/epjds/s13688-018-0163-7

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Big data would not lie: prediction of the 2016 Taiwan election via online heterogeneous information

Zheng Xie[1], Guannan Liu[2*], Junjie Wu[2,3] and Yong Tan[4]

*Correspondence: liugn@buaa.edu.cn
[2]School of Economics and Management, Beihang University, Beijing, China
Full list of author information is available at the end of the article

Abstract

The prevalence of online media has attracted researchers from various domains to explore human behavior and make interesting predictions. In this research, we leverage heterogeneous data collected from various online platforms to predict Taiwan's 2016 general election. In contrast to most existing research, we take a "signal" view of heterogeneous information and adopt the Kalman filter to fuse multiple signals into daily vote predictions for the candidates. We also consider events that influenced the election in a quantitative manner based on the so-called event study model that originated in the field of financial research. We obtained the following interesting findings. First, public opinions in online media dominate traditional polls in Taiwan election prediction in terms of both predictive power and timeliness. But offline polls can still function on alleviating the sample bias of online opinions. Second, although online signals converge as election day approaches, the simple Facebook "Like" is consistently the strongest indicator of the election result. Third, most influential events have a strong connection to cross-strait relations, and the Chou Tzu-yu flag incident followed by the apology video one day before the election increased the vote share of Tsai Ing-Wen by 3.66%. This research justifies the predictive power of online media in politics and the advantages of information fusion. The combined use of the Kalman filter and the event study method contributes to the data-driven political analytics paradigm for both prediction and attribution purposes.

Keywords: Election prediction; Heterogeneous data; Kalman filter; Event study method; Big data

1 Introduction

Recent years have witnessed the rapid development of social media and their innovative applications in many fields [1]. For instance, it has been found that the volumes of tweets related to protests on Twitter are associated with real-life protest events [2]. Moreover, film mentions on Twitter can reflect box office revenues [1]. Additionally, public moods extracted from tweets can predict changes in stock markets [3, 4], and a real-time earthquake reporting system was developed by analyzing only tweets [5].

The unprecedented prevalence of social media has driven politicians to make use of this channel to propagate their ideas and political views [6–9] to more directly approach potential voters. It is not unusual to see election candidates post their daily activities and political ideas on social media and even debate on social media before and during the campaign. These behaviors can attract online discussion from massive numbers of netizens and, compared with traditional polls, are an easier way to gather wide-ranging public opinions about the candidates. Some research has shown the predictability of election results based on social media information in various countries and regions, including the United States [10–12], the United Kingdom [13], Germany [14], the Netherlands [15], and Korea [16], where netizens' behaviors and posts on social media were analyzed to infer the election results.

The existing research, however, usually exploits a single information source and uses simple descriptive statistics for election predictions, which easily results in hindsight bias and lacks generality. The way to ameliorate these issues is two-fold. On one hand, multiple sources should be included to obtain heterogeneous information for robust predictions. For instance, the keywords searched in Google represent the attention of the public, and the aggregated volumes can be used to predict the trends of influenza [17], stock markets [18, 19], consumer behaviors [20], *etc*. On the other, massive heterogeneous data obtained in real time are often too chaotic to provide consistent predictions; therefore, a method that can fuse the data and deliver robust predictions is indispensable. Our work in this paper is a novel attempt on this front.

We take Taiwan's 2016 general election as a real-life case. Taiwan adopted direct election in 1996, and since then, Kuomintang (KMT) and the Democratic Progressive Party (DPP) have become the two major competing political parties. KMT pursues a "One China Policy" and the political legitimacy of the "Republic of China", whereas DPP takes "Taiwan Independence" as its party program. In 2016, three candidates ran for the general election, including Eric Chu from KMT, Tsai Ing-wen from DPP, and James Soong from the People First Party (PFP). The election regulations adopt the "one man one vote" principle and execute the majority rule [21].

This research leverages time series data collected from various mainstream online platforms (*i.e.*, Facebook, Twitter and Google) and visitation traffic to candidates' campaign pages. These heterogeneous signals represent public opinions and are fed into a *Kalman filter* [22] to estimate the vote shares of each candidate dynamically. The most efficient signals are then identified based on the signal strengths characterized by the *Kalman gain*. In addition to prediction, this research attempts to automatically identify the events that most influenced the election by leveraging the *event study model* that originated in the field of financial research [23].

The results show that the prediction errors for every candidate one day, week, and month before the election are no greater than 2.59%, 4.58% and 5.87%, respectively. The results include some interesting findings. First, online signals appear to be more accurate than traditional polls in election prediction, although the polls can still function on mitigating the sample bias of netizens. In particular, a simple Facebook "Like" on a candidate's post is the most significant predictor, whereas the seemingly more informative "Comments" function is much less important. Second, online signals show clear convergence as the final election day approaches. For example, Google keyword searches fluctuated initially but became a strong indicator in the final stage. Third, bursty events most influential to the campaign have a strong relationship with the cross-strait relation topics. For instance, while the Xi-Ma meeting reduced support of Tsai Ing-wen by 0.55%, the Chou

Tzu-yu flag incident followed by the apology video one day before the election increased her votes by 3.66%.

2 Data and measurements

To identify the most popular Internet applications in Taiwan, we referred to professional Internet surveys[a] and web traffic reports from Alexa, comScore and Digital Age (see Additional file 1, Table S1). We selected Facebook, Twitter, Google, and candidates' campaign homepages as the "online sensors" of public opinions towards the election and designed various daily updated measurements to characterize the signals during the period from Oct. 31, 2015 to Jan. 16, 2016 consecutively. A 30-day moving average was applied to each measure to avoid excessive fluctuation. The data sets are available from: https://doi.org/10.6084/m9.figshare.6014159.

Facebook. Facebook is the most popular social platform in Taiwan and provides an easy way for candidates to reach out to a large audience. For each post by a candidate, users can click the "Like" tag to indicate a positive reaction. Hence, we can use the "daily average number of Likes per post" to measure a candidate's popularity:

$$s^{c}_{k,\mathrm{FAL}} = \frac{1}{m}\sum_{j=0}^{m-1}\frac{\sum_i like^{c}_{k-j,i}/n^{c}_{k-j,\mathrm{FA}}}{\sum_c\sum_i like^{c}_{k-j,i}/n^{c}_{k-j,\mathrm{FA}}}, \tag{1}$$

where $like^{c}_{k,i}$ is the number of *Likes* of post i published by candidate c on day k, $n^{c}_{k,\mathrm{FA}}$ is the total number of the candidate's posts, and m is the window length of the moving average. Analogously, we compute the "daily average number of Comments per post" for each candidate as another signal from Facebook:

$$s^{c}_{k,FAC} = \frac{1}{m}\sum_{j=0}^{m-1}\frac{\sum_i Comment^{c}_{k-j,i}/n^{c}_{k-j,\mathrm{FA}}}{\sum_c\sum_i Comment^{c}_{k-j,i}/n^{c}_{k-j,\mathrm{FA}}}, \tag{2}$$

where $Comment^{c}_{k,i}$ is the number of comments on post i published by candidate c on day k.

Twitter. We use three candidates' names in both Simplified and Traditional Chinese as keywords (see Additional file 1, Table S2) to retrieve tweets from Twitter. The measure "number of tweets mentioning the candidate" is calculated as

$$s^{c}_{k,\mathrm{TW}} = \frac{1}{m}\sum_{j=0}^{m-1}\frac{tw^{c}_{k-j}}{\sum_c tw^{c}_{k-j}}, \tag{3}$$

where tw^{c}_{k} is the volume of tweets about candidate c on day k.

Search Engine. We also obtained search data from Google Trends to trace the evolution of a keyword's search volume. We used the three candidates' names in both Simplified and Traditional Chinese as keywords and restricted the search source to Taiwan. The measurement "search index ratio" is defined as

$$s^{c}_{k,\mathrm{GO}} = \frac{1}{m}\sum_{j=0}^{m-1}\frac{search^{c}_{k-j}}{\sum_c search^{c}_{k-j}}, \tag{4}$$

where $search^{c}_{k}$ is the aggregated search indexes of keywords about candidate c on day k.

Campaign Homepages. We collected the daily traffic to candidates' campaign homepages data from Alexa, and used the "IP traffic ratio" as an opinion measure as follows:

$$s^c_{k,\mathrm{IP}} = \frac{1}{m}\sum_{j=0}^{m-1}\frac{\mathrm{IP}^c_{k-j}}{\sum_c \mathrm{IP}^c_{k-j}}, \tag{5}$$

where IP^c_k is the IP traffic volume to candidate c's campaign homepage on day k.

The above measurements convey different signals for continuous election prediction. We also collected offline election polls published by nineteen authoritative pollsters during the period from Aug. 1, 2015 to Jan. 16, 2016 (see Additional file 1, Sect. 1.1) for comparison. These polls were published aperiodically and infrequently, so we assume the opinions from a poll remain unchanged until a new poll has been released.

3 Vote prediction model

The goal of election prediction is to infer the underlying vote shares of various candidates based on heterogeneous noisy signals. A model that can fuse the signals in such a way to debias the prediction from noise and make dynamic predictions to reflect the evolution of public opinion is desired. We exploit the *Kalman filter*, a linear dynamic model, for this purpose. The filter was adopted in [24–26] for election analysis, but previous studies were mostly based on polls and assumed only two candidates.

In general, a Kalman filter maps hidden states to observed variables with noise, and the current hidden states are assumed to transition from previous states with noise. That is,

$$\begin{aligned} \mathbf{s}^c_k &= \mathbf{h}_k x^c_k + \mathbf{r}^c_k, \quad \mathbf{r}^c_k \sim N\bigl(0, \mathbf{R}^c_k\bigr), \\ x^c_k &= f_k x^c_{k-1} + q^c_k, \quad q^c_k \sim N\bigl(0, \sigma^2_{c,k}\bigr), \\ x^c_0 &\sim N\bigl(m^c_0, p^c_0\bigr), \end{aligned} \tag{6}$$

where $\mathbf{h}_k$ is a vector that maps the hidden state x^c_k of candidate c to observed multiple signals in $\mathbf{s}^c_k$, f_k is the state transition coefficient, and x^c_0 is the initial value of the hidden state. $\mathbf{r}^c_k$ and q^c_k denote independent Gaussian random noise.

In our case, x^c_k is the genuine vote share of candidate c on day k, and $\mathbf{s}^c_k = (s^c_{k,\mathrm{GO}}, s^c_{k,\mathrm{FAL}}, s^c_{k,\mathrm{TW}}, s^c_{k,\mathrm{IP}})^\top$ contains the observed multiple signals. We set $f_k = 1$ and $\mathbf{h}_k = \mathbf{1}$ for scale equivalence of the variables. The initial vote m^c_0 is set as the average value of the latest poll results, with $p^c_0 = 1$ to allow fluctuation. Note that we also change the setting of initial vote m^c_0 to the mean value of each candidates' signals and an equal value $m^c_0 = 1/3$, with state variances $p^c_0 = 0$ and $p^c_0 = 1$, respectively (see Additional file 1, Sect. 2.1). The final prediction turns out to be insensitive to the initial values when the time series is sufficiently long (see Additional file 1, Sect. 2.2 and Sect. 2.3). The logic behind the set of equations is that the online measures are flawed signals with the true vote states represented by the mean with mixing noise. The goal of the model is to fuse the flawed signals to estimate the daily state and to further transfer the estimation to the next day to make a prediction.

The next task is to estimate the noise parameters $\mathbf{R}^c_k$ and $\sigma^2_{c,k}$. To reduce the model complexity, we assume $\mathbf{R}^c_k = \mathbf{R}_k$ and $\sigma^2_{c,k} = \sigma^2_k$, $\forall c$. The maximum a posteriori estimation can then be obtained by maximizing the conditional density function:

$$\begin{aligned}\mathcal{J} &= p\big(x_{1:k}^{tsai}, x_{1:k}^{chu}, x_{1:k}^{soong}, \sigma_k^2, \mathbf{R}_k | \mathbf{s}_{1:k}^{tsai}, \mathbf{s}_{1:k}^{chu}, \mathbf{s}_{1:k}^{soong}\big) \\ &\propto \prod_c p\big(x_0^c\big) \prod_{j=1}^{k} p\big(\mathbf{s}_j^c | x_j^c, \mathbf{R}_k\big) p\big(x_j^c | x_{j-1}^c, \sigma_k^2\big) p\big(\sigma_k^2, \mathbf{R}_k\big),\end{aligned} \tag{7}$$

with $\sum_c x_k^c = 1$ and $\sum_c \mathbf{s}_k^c = \mathbf{I}_{4\times 1}$. We finally have (see Additional file 1, Sect. 2.1),

$$\begin{aligned}\widehat{\sigma_k^2} &= \frac{1}{3k}\sum_c \sum_{j=1}^{k} \big(\hat{x}_{j|j}^c - f_j \hat{x}_{j-1|j-1}^c\big)^2, \\ \widehat{\mathbf{R}}_k &= \frac{1}{3k}\sum_c \sum_{j=1}^{k} \big(\big(\mathbf{s}_j^c - \mathbf{h}_j \hat{x}_{j|j-1}^c\big)\big(\mathbf{s}_j^c - \mathbf{h}_j \hat{x}_{j|j-1}^c\big)^\top - \mathbf{h_j} p_{j|j-1}^c \mathbf{h_j}^\top\big),\end{aligned} \tag{8}$$

where $\hat{x}_{k|k-1}^c$ is the vote state prediction for candidate c at time k given the signals up to $k-1$, and $\hat{x}_{k|k}^c$ is the updated estimation of the vote state at time k given the signals up to k. $p_{k|k-1}^c$ and $p_{k|k}^c$ are the prediction covariance and updated estimation covariance, respectively.

To recursively estimate the daily vote state at time k, the prediction of vote shares $\hat{x}_{k|k-1}^c$ is first derived by a variation of the state transition equation in (6):

$$\begin{aligned}\hat{x}_{k|k-1}^c &= f_k \hat{x}_{k-1|k-1}^c, \\ p_{k|k-1}^c &= f_k^2 p_{k-1|k-1}^c + \widehat{\sigma_k^2}.\end{aligned} \tag{9}$$

Meanwhile, since the online signal $\mathbf{s}_k^c$ is observed, it is feasible to update the state estimation $\hat{x}_{k|k}^c$ by absorbing $\mathbf{s}_k^c$ into the prediction of $\hat{x}_{k|k-1}^c$. We use a weighted function to express the combination of the state prediction and signals as follows:

$$\begin{aligned}\hat{x}_{k|k}^c &= f_k \hat{x}_{k|k-1}^c + \mathbf{k}_k^c\big(\mathbf{s}_k^c - \mathbf{h}_k \hat{x}_{k|k-1}^c\big), \\ p_{k|k}^c &= p_{k|k-1}^c - \mathbf{k}_k^c \mathbf{h}_k p_{k|k-1}^c,\end{aligned} \tag{10}$$

where $\mathbf{k}_k^c$ is called the *Kalman gain* [27] used to weight the state prediction and various signals in the prediction. By minimizing the updated state estimation error $x_k^c - \hat{x}_{k|k}^c$, we can derive the Kalman gain as

$$\mathbf{k}_k^c = p_{k|k-1}^c \mathbf{h}_k^\top \big(\mathbf{h}_k p_{k|k-1}^c \mathbf{h}_k^\top + \widehat{\mathbf{R}}_k^c\big)^{-1}. \tag{11}$$

When the updated estimation is obtained, we can use (9) to predict the next-day vote share.

According to the Internet usage report of Taiwan,[a] more than 90% of Taiwan residents aged between 20 and 45 years have accessed the Internet since May 2015. This proportion is over 80% in the population aged between 45 and 55 years. By contrast, only 49.5% of residents aged over 55 years have used the Internet during the same time period. Thus, we take the online data fusion result as a representation for the group aged between 20 and 50 years. With respect to the age-adjusted sampling method adopted by pollsters, we take the poll results for the 50 to 60 year-old, 60 to 70 year-old and over 70 year-old groups as the vote share estimations of the corresponding age groups. Therefore, the final daily vote share prediction y_k^c for candidate c at time k is weighted as follows,

$$y_k^c = w_{20\sim 50}\hat{x}_{k|k-1}^c + w_{50\sim 60}z_{50\sim 60,k}^c + w_{60\sim 70}z_{60\sim 70,k}^c + w_{70}z_{>70,k}^c, \tag{12}$$

where w_i is the population proportion of age group i, which could be obtained from the Ministry of the Interior of Taiwan.[b] $z_{i,k}^c$ is the most recent poll result of age group i for candidate c on day k.

4 Event detection method

Twitter, as an online plaza, aggregates information about different candidates during an election campaign. By analyzing the sentiment of Twitter in October 2015, we find that more than 80% of the retrieved tweets are news. Due to the fact that most of the Taiwan mainstream media have set up accounts in Twitter, the volatility of tweets is able to signal influential events. A three-step detection method is designed as follows.

Step I is to perceive events based on massive numbers of tweets. To this end, we watch the statistic tw_k^c, *i.e.*, the number of tweets about candidate c on day k, and trace its volatility in the past m days by comparing it with an upper bound $u_{k+1}^c = \bar{n} + \frac{s}{\sqrt{m}}t_{\alpha/2}(m-1)$, where $\bar{n}$ is the average of tw_k^c on m days and s is the standard deviation. Based on a t-test with significance level α, there exists an influential event if tw_{k+1}^c surpasses u_{k+1}^c (see Additional file 1, Fig. S9). We assume that only one new event is dominant in each burst, which is reasonable for political campaigns.

Step II is to estimate the event time window. The daily tweets about each candidate are first integrated into a single document; then, the terms in the document are weighted by the *tf-idf* method. *tf-idf* is a numerical statistic intended to reflect how important a word is to a document in a collection of corpora. The *tf-idf* value increases proportionally with the number of times a word appears in a document but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. *tf-idf* is calculated as follows,

$$\begin{aligned} tf\left(t, d_k^c\right) &= \frac{f_{t,d_k^c}}{\sum_t f_{t,d_k^c}}, \\ idf\left(t, D^c\right) &= \log \frac{N^c}{1 + |d_k^c \in D^c : t \in d_k^c|}, \\ tf\text{-}idf\left(t, d_k^c, D^c\right) &= tf\left(t, d_k^c\right) idf\left(t, D^c\right), \end{aligned} \tag{13}$$

where f_{t,d_k^c} is the count of term t in a tweet d_k^c referring to candidate c on day k. D^c is the total tweets of candidate c, $N^c = |D^c|$, and $|d_k^c \in D^c : t \in d_k^c|$ is the number of documents in which the term t appears. The top-30 terms with the highest weights in the burst are selected as the typical words for that event. We then proceed to check the overlaps of typical words on the burst day plus or minus five days. The first day with non-zero overlap is deemed to be the start day of the event, and the last day with non-zero overlap is the closing day, which defines the event time window (see Additional file 1, Table S9, Table S10, and Table S11). We remove suspicious events with a time window of only one day.

Step III is to measure the impact of events on public opinion. We denote the estimated x^c_k initially transited from the previous day as $\hat{x}^c_{k|k-1}$ (see equation (9)) and the final x^c_k calibrated with multiple signals as $\hat{x}^c_{k|k}$ (see equation (10)). Intuitively, $\hat{x}^c_{k|k}$ has absorbed the information about all pertinent events on day k; hence, the change from $\hat{x}^c_{k|k-1}$ (equaling $\hat{x}^c_{k-1|k-1}$ for $f_k = 1$ and $\mathbb{E}(q^c_k) = 0$) to $\hat{x}^c_{k|k}$ indicates the impact of an event. To measure the significance of the impact, we apply the *event study model* [28] from the field of finance as follows:

$$\hat{x}^c_{k|k} = a + \hat{x}^c_{k-1|k-1} + \sum_{j=1}^{J} \gamma_j D^c_{j,k} + \varepsilon, \tag{14}$$

where $D^c_{j,k}$ is a dummy variable equal to 1 if day k is within the time window of event j for candidate c and is equal to 0 otherwise. J is the total number of detected events, and a is a regression constant. γ_j is the estimator of the effect of event j, which passes the t-test if event j has a significant effect on public opinion. In this way, we can identify the events that actually influence the election.

5 Results

5.1 Prediction performance

Figures 1(a)–(c) show various online signals two months before election day. Intuitively, the user behavior in different channels is related to the public opinion towards a candidate, but the signals have vastly different volatilities. This justifies the value of information fusion for election prediction.

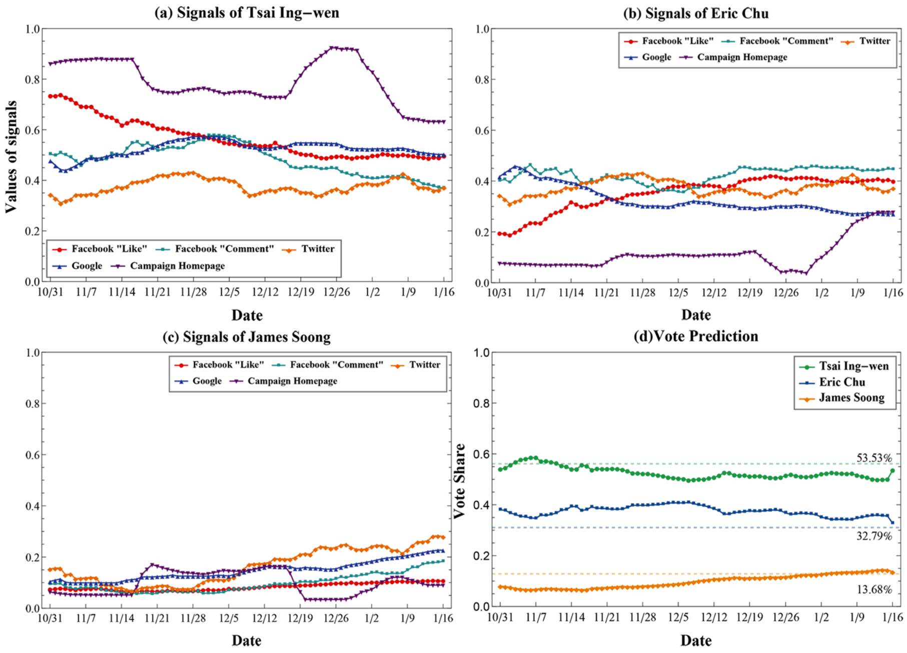

Figure 1 Online signals and time series of vote share predictions. (**a**)–(**c**) Signals of public opinions from all the online channels for the three candidates. (**d**) Time series of the vote share predictions. The dashed lines are the actual election outcomes. On election day, the errors are less than 2.59%

Figure 1(d) depicts the dynamic vote predictions after fusing the four types of online signals, *i.e.*, $s^c_{k,\mathrm{FAL}}$, $s^c_{k,\mathrm{TW}}$, $s^c_{k,\mathrm{GO}}$ and $s^c_{k,\mathrm{IP}}$, by the Kalman filter. Although the four signals behave differently, the fused signal representing the predicted vote share for each candidate is relatively stable and exhibits a clear tendency, confirming the effectiveness of the prediction system for information aggregation. The final result is impressive—while Tsai's win is easy to predict even in October, the prediction errors for every candidate one day, week, and month before the election day are no greater than 2.59%, 4.58% and 5.87%, respectively.

To further justify the predictive power of online signals, we also compare our results with offline polls. As shown in Fig. 2, during the last two weeks of the election, our predictions (M1) outperform most of the pollsters (P1–P10), and can improve continuously by absorbing up-to-date information. This is possibly due to the fact that the anonymity of the Internet enables individuals to express their opinions freely and voluntarily, which could reduce the bias relative to that in the tele-interview setting of a traditional poll. Furthermore, currently, news usually breaks online first and then spreads at a tremendously fast pace from online to offline via physical social networks. Therefore, online information can also influence offline voting blocs during campaigns, which mitigates the bias effect of using only the netizen population in our method.

We also try to reduce the sample bias by mixing the prediction results from online signals with those from offline pollsters in older groups. As shown in Fig. 2, the online-offline data fusion method (M2) indeed outperforms the online data fusion method (M1) in the early stage of the final two weeks, which indicates the power of sample bias correction. But the advantage disappears gradually as the final election day approaches, which again exposes the drawback of offline polls in responding to newly emerging information.

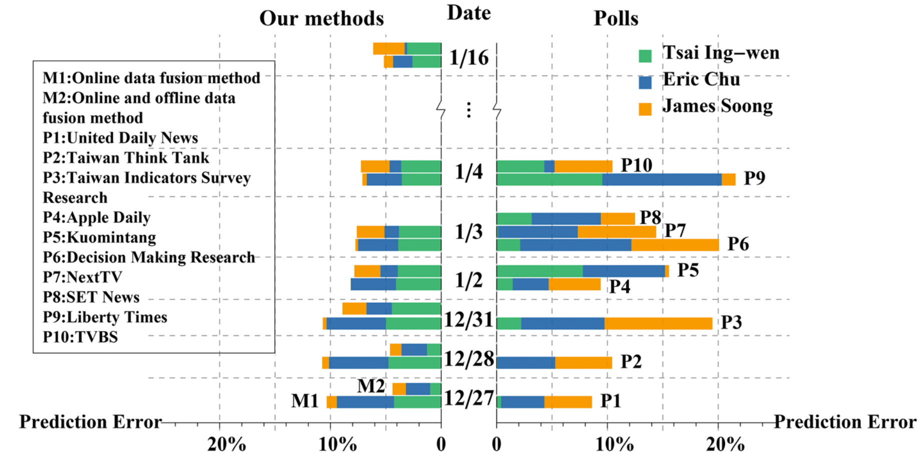

Figure 2 Timeline of the absolute prediction errors of final polls and data fusion methods. The bars on the left side of the timeline represent the prediction errors of the data fusion methods. In each interval between two gray dashed lines, there are two bars. The lower bar represents the absolute error of the online data fusion method, and the upper bar represents the absolute error of the online–offline data fusion method. The interval between two gray horizontal dashed lines indicates one day. The bars on the right side of the timeline show the prediction errors of the final polls from ten pollsters. Comparison of the bars on both sides shows that the absolute prediction errors of the signal fusion methods are smaller than those of the polls

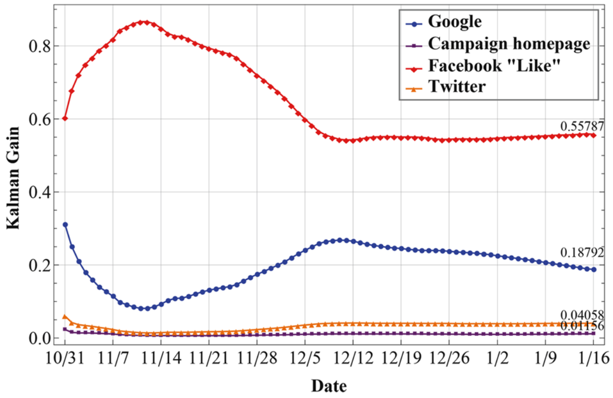

Figure 3 Kalman gain for different online signals. The Kalman gain of the Facebook "Like" constitutes the highest proportion but gradually decreases while the Google signal continuously increases. Twitter and the campaign homepages are not indicative signals for the election. The Kalman gains of all the signals converge to steady states approximately one month before the election

5.2 Signal evaluation

We also explore the predictive power of various online signals via their daily Kalman gains $\mathbf{k}_k^c$. As shown in Fig. 3, Facebook "Likes" are consistently the strongest indicator among all the signals. This demonstrates the power of social media in collecting public opinions via a simple mechanism, although it is vulnerable to shilling attacks. The predictive power of the Google index appears to be time-sensitive, contributing less initially and becoming the second best indicator one month before the election. One possible explanation is that the election might not be a focal topic in the early stage of the campaign, making Google searches rather random. However, as the election day approaches, the campaign becomes the central topic and drives the public to search for information about the candidates. The two remaining signals, *i.e.*, tweet volumes and homepage traffic, appear to be of much weaker predictive value, which may be due to their lack of popularity in Taiwan (see Additional file 1, Table S1) and diverse attitudes about candidates.

We further explore the distinct value of the "Like" function on Facebook. We compare it with the "Comment" function by substituting $s_{k,\mathrm{FAL}}^c$ with $s_{k,FAC}^c$ in the Kalman filter. The results indicate that the prediction outcomes become significantly worse—the one-day-earlier prediction errors for Tsai and Chu increase to 5.42% and 4.86%, respectively (see Additional file 1, Sect. 2.5). These results indicate the superiority of "Like" over "Comment". To understand this result, we search for the population of Facebook users who have ever liked or commented on the candidates and obtain the overlapping users who have both liked and commented on a candidate. Figure 4 shows that these users constitute only a small proportion of the "Like" users but a much larger proportion of the "Comment" ones. Therefore, a considerable proportion of users who have commented on a post may also choose to like the post but not vice versa. In other words, the "Like" signal represents the positive attitude of a much larger population than that of the "Comment" signal, which may be attributed to the fact that a "Like" is a more direct and widely engaged in behavior for online users to express their positive opinions without great effort. Another disadvantage of "Comment" lies in its diversity of expression, which can be a blend of contradictory attitudes, including support, praise, opposition and even insult. We apply Latent Dirichlet Allocation (LDA) model [29] to extract topics from the overlapping users and users who

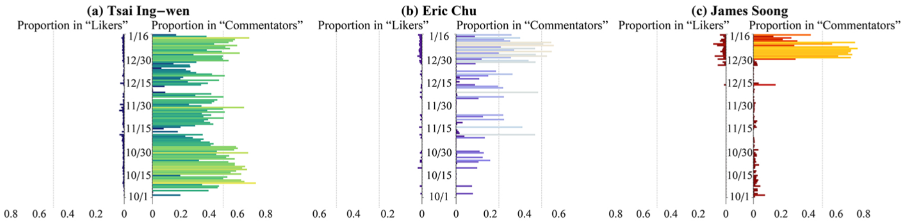

Figure 4 Proportions of overlapping users who have both liked and commented on the candidates' posts among "Likers" and "Commentators". The intermediate vertical axis is a timeline covering the whole period of the election. The bar on the left side of the timeline represents the daily proportion of overlapping users to users who have ever liked. The bar on the right side of the timeline represents the daily proportion of overlapping users to users who have ever commented. The number of overlapping users accounts for less than 1% of all the users who have "liked" on average, with the maximum proportions being 3.51%, 3.74%, and 9.25% for the three candidates. By contrast, the overlapping users constitute more than 37.16%, 14.90%, and 12.03% of all users who have commented, on average, for the three candidates, and the maximum ratios are 73.05%, 59.75%, and 83.01%

only commented on the candidates. The representative topics of the overlapping users are mainly supportive attitudes, while the topics of the users who only commented on candidates are mixed, with both positive and negative topics (see Additional file 1, Tables S3–S8).

The overlapping users indeed constitute a group of firm supporters for each candidate who show their support by not only clicking "Like" but also going through the effort to publish comments. By further tracking the changes in the overlap ratios during the election, as shown in Fig. 4, we find that the ratio for Tsai is relatively stable, indicating that Tsai has a firm group of supporters regardless of her behavior during the campaign. By contrast, for Chu and Soong, the overlap ratios remain small until election day approaches, suggesting Tsai should partially attribute her success to her firm supporters rather than swing voters. This also explains why we can predict the victory of Tsai two months before election day.

5.3 Influential events

We apply the event detection method to each candidate's Twitter data to identify influential events. Figure 5 shows the results, and Table 1 shows the event descriptions. The most influential events detected with p-values less than 0.05 include the meeting between Xi Jinping and Ma Ying-jeou (Xi-Ma Meeting), the emergence of negative comments on Tsai Ing-wen's Facebook homepage possibly by users from mainland China, and the Chou Tzu-yu flag incident. All these events share a common feature; that is, they all belong to the category of *cross-strait relation*, which is always subtle and controversial in Taiwan's political circle. Other seemingly important events from the perspective of the election campaign, such as the TV broadcast of the candidates' debates and various types of electioneering activities in local areas, have insignificant influences on public opinion.

We further assess the influence level of the events, which is measured by the coefficient γ_j in (14). Table 2, Table 3 and Table 4 give the detailed results for the three candidates, respectively. The statistical results of γ_i, $i \in \{1,\ldots,21\}$, correspond to the effects of 21 events marked in E_i, $i \in \{1,\ldots,21\}$, in Table 1.

The Xi-Ma Meeting resulted in a 0.55% decrease in the vote share of Tsai Ing-wen. This result is not surprising because Tsai was believed to favor Taiwan independence over the

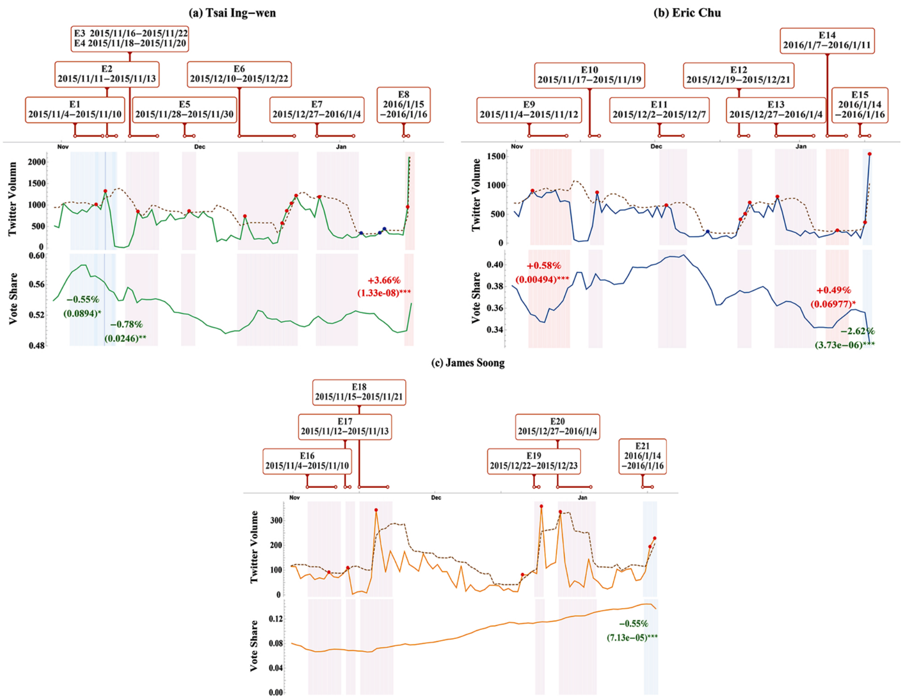

Figure 5 Detected events and their influence on potential vote shares. The numbers of detected events during the elections are 8, 7, and 6, respectively, for the Tsai, Chu, and Soong. The description of events E_i is presented in Table 1. Each event spans a time window and influences the potential voting rates differently, as denoted by different colors. A light purple bar indicates that the detected event does not have a significant influence on the vote shares. A red bar indicates a positive effect of the event on vote shares, and a blue bar represents a negative effect of the event. The influence of each significant event is marked on the curve, and the number below it in brackets is the *p*-value of the t-test. The typical words used to determine the event timespan are detailed in Additional file 1, Table S9, Table S10, and Table S11. In addition, the Twitter bursty days detected in Step I are noted in the Twitter volume time series. The red points represent events with a timespan longer than one day, which are fed into Step II for further analysis. The blue points are removed

"One China Policy", and the meeting thus prompted the public to doubt Tsai's ability to handle cross-strait relations. This same event increased Eric Chu's vote share by 0.58% because he was thought to be more able to develop cross-strait peace after the meeting.

Despite the abundance of events during the campaign, the *Chou Tzu-yu flag incident* from the entertainment domain is the most influential. Chou Tzu-yu, a 16-year-old Taiwan singer, sparked huge controversy in social media for showing the Taiwan flag as the national flag of China. As the uproar intensified online, Chou's company released a video in which Chou apologized for her behavior by stating that "there is only one China" and identifying herself as Chinese. The most subtle point is that the video was released the day before the election, which was described as a humiliation to Taiwan and spread quickly in Taiwan's online social media. As a consequence, this incident increased the vote share of Tsai Ing-wen by approximately 3.66% and lowered the vote share of Eric Chu by approximately 2.62%.

6 Discussion

The accurate prediction of Taiwan's 2016 general election suggests an interesting viewpoint that public opinions towards political campaigns can be determined via online user-generated content. This indeed coincides with some recent studies reporting that social

Table 1 Detected Events

Tsai Ing-wen		Eric Chu		James Soong	
Mark	Event	Mark	Event	Mark	Event
E1	Xi-Ma Meeting	E9	Xi-Ma Meeting	E16	Xi-Ma Meeting
E2	Negative comments written in Simplified Chinese spam Tsai Ing-wen's Facebook homepage	E10	Eric Chu announces Wang Ju-hsuan as his running mate	E17	James Soong takes part in the commemoration of Sun-Yat-sen
E3	Tsai Ing-wen announces Chen Chien-jen as her running mate	E11	Eric Chu names Hu Chih-chiang the manager of his campaign	E18	James Soong announces Hsu Hsin-ying as his running mate
E4	Tsai Ing-wen becomes the cover star of the Economist	E12	Kao Wan-ching, Eric Chu's wife, canvasses for him	E19	James Soong attends the business forum
E5	The verdict of Ting Hsin International Group arouses the strong dissatisfaction of the public	E13	Television debates	E20	Television debates
E6	Tsai Ing-wen is accused of speculation on land	E14	Eric Chu proposes that the party holding the majority in the new Legislature is allowed to form a new cabinet	E21	Chou Tzu-yu flag incident
E7	Television debates	E15	Chou Tzu-yu flag incident		
E8	Chou Tzu-yu flag incident				

*The same event can have different impacts on the three candidates and thus is labeled differently.

Table 2 Influential significance of events detected for Tsai Ing-wen

	Estimate	Std. Error	*t* value	Pr(>|*t*|)
c	−0.0262	0.0224	−1.169	0.2464
p_{k-1}^{tsai}	1.0503	0.0430	24.414	<2e−16***
γ_1	−0.0055	0.0032	−1.724	0.0894*
γ_2	−0.0078	0.0034	−2.299	0.0246**
γ_3	−0.0016	0.0027	−0.611	0.5430
γ_4	0.0007	0.0039	0.190	0.8499
γ_5	−0.0024	0.0029	−0.839	0.4047
γ_6	0.0011	0.0019	0.604	0.5480
γ_7	0.0007	0.0021	0.362	0.7188
γ_8	0.0366	0.0057	6.469	1.33e−08***
Residual standard error	0.0055 on 67 degrees of freedom			
Multiple R-squared	0.9529			
Adjusted R-squared	0.9459			
F-statistic	135.6 on 10 and 67 DF	*p*-value	<2.2e−16	

***, **, and * indicate that the coefficients are significantly different from zero at the level of 1%, 5%, and 10%, respectively.

media such as Facebook [6, 10], Twitter [2, 6, 7, 11, 13–16] and Youtube [6] are able to aggregate public opinions about political matters. Donald Trump winning the 2016 US Presidential Election was also considered to be a victory for the heavy use of social media such as Twitter [30]. Nevertheless, this finding remains controversial in academia, and the above studies have often been criticized for the unreliability of single-source information [31] and/or the unrepresentativeness of online user populations [32, 33]. Our study attempts to address these concerns.

First, we introduce multiple online channels as different types of signals to produce more robust predictions. These signals, while reflecting more or less latent public opinions, have varied fluctuations due to their different sensitivities to campaign dynamics and possible

Table 3 Influential significance of events detected for Eric Chu

	Estimate	Std. Error	*t* value	Pr(>|*t*|)
c	–0.0056	0.0183	–0.307	0.76012
p_{k-1}^{llchu}	1.0102	0.0481	21.004	<2e–16***
γ_9	0.0058	0.0020	2.906	0.0049***
γ_{10}	0.0025	0.0027	0.945	0.3482
γ_{11}	0.0010	0.0024	0.418	0.6776
γ_{12}	0.0023	0.0026	0.871	0.3870
γ_{13}	–0.0002	0.0021	–0.110	0.9130
γ_{14}	0.0049	0.0026	1.842	0.0698*
γ_{15}	–0.0262	0.0052	–5.036	3.73e–06***
Residual standard error	0.0050 on 68 degrees of freedom			
Multiple R-squared	0.9529			
Adjusted R-squared	0.9466			
F-statistic	152.7 on 9 and 68 DF	*p*-value	<2.2e–16	

***, **, and * indicate that the coefficients are significantly different from zero at the level of 1%, 5%, and 10%, respectively.

Table 4 Influential significance of events detected for James Soong

	Estimate	Std. Error	*t* value	Pr(>|*t*|)
c	–0.0009	0.0012	–0.735	0.4651
p_{k-1}^{llchu}	1.0172	0.0114	89.259	<2e–16***
γ_{16}	–0.0004	0.0008	–0.488	0.6273
γ_{17}	–0.0006	0.0011	–0.590	0.5572
γ_{18}	0.0009	0.0007	1.241	0.2192
γ_{19}	–0.0006	0.0010	–0.560	0.5774
γ_{20}	0.0004	0.0007	0.550	0.5844
γ_{21}	–0.0055	0.0013	–4.227	7.13e–05***
Residual standard error	0.0017 on 69 degrees of freedom			
Multiple R-squared	0.9963			
Adjusted R-squared	0.9959			
F-statistic	2331 on 8 and 69 DF	*p*-value	<2.2e–16	

***, **, and * indicate that the coefficients are significantly different from zero at the level of 1%, 5%, and 10%, respectively.

fake responses from the Internet "water army" (see Fig. 1). The fusion of these signals can help to filter out some noise by consensus learning to highlight the tendencies. Moreover, although one signal might contribute more to some specific election prediction, such as the Facebook "Like" for the Taiwan election, it is unlikely to find it omnipotent for different elections. The fusion of these signals could help to mitigate the risk of selection bias. This information fusion scheme gives our study some important extensibility—the four channels, namely, Facebook, Twitter, Google Trends and campaign homepages, could be considered to be the fundamental and preemptive online information sources for different elections.

We also find that although selection bias of the online voting population exists, its influence on the prediction results is limited. Prediction based on pure online information is much more accurate than the polls released by Taiwan's mainstream pollsters (see Fig. 2). The reason behind this may be two-fold. On one hand, online users who pay close attention to election campaigns likely become active voters and constitute a large voting population on election day [34, 35]. On the other hand, we should not underestimate the information exchange between online social networks and offline physical networks [36, 37]. Older people who seldom interact with the Internet still have access to online information via ordinary family communications or traditional media's reports on Internet opinions. This communication contributes to the opinion conformance across online and offline networks and further improves the representativeness of the online voting population. In fact, compared with traditional polls, which are susceptible to questionnaire wording [38], reporting error [39], ballot order [40], and social desirability bias [39, 41], online big data enables a much larger sample and thus can improve the sample resistance to human manipulation. The real-time availability of online data, which enables dynamic predictions based on continuously incoming information, is another major advantage relative to polls.

Our study also suggests that the Kalman filter with the event detection model could be packaged as a fundamental kit for political vote analytics. Specifically, the Kalman filter is responsible for the dynamic prediction of vote shares given multi-source time-varying signals and multiple candidates. Meanwhile, the event detection model is responsible for the automatic identification of influential events during the campaign, which provides a causal explanation for the predictions. In other words, the two models together could provide *interpretable predictions* to political vote analytics, which is deemed particularly valuable for a big-data-driven research paradigm [42].

The Kalman filter has been adopted in previous studies but either for backward review given the final result or for forward prediction given multiple historical elections data. Our study shows that while we cannot obtain the true vote shares until election day, we can still fine-tune the model parameters by using up-to-date time series signal data for the current election, which solves the problems in leveraging the Kalman filter for election prediction. Moreover, given the sum-to-one constraint in a statistical learning framework (see (8)), the Kalman filter is capable of building models for more than two election candidates. One may consider the inclusion of some other relatively stable factors, such as the globalization trend, economic status, the technology environment, *etc.*, in the prediction model, which can be achieved by setting appropriate initial values of the Kalman filter. Nevertheless, our study shows that the Kalman filter is insensitive to the initial values as long as the prediction is based on a sufficiently long time series (see Additional file 1, Sect. 2.2 and Sect. 2.3). In this case, the signals should have fully "absorbed" the influences of the macro factors.

Our study provides some political insight into the Taiwan general election. It is interesting that the simple "Like" function on Facebook collects the public opinions about candidates (see *Signal Evaluation* in *Results*), although it has been reported to be vulnerable to shilling attacks in electronic commerce [43]. The "Like" function is more beneficial than the "Comment" function, although the latter actually expresses more complex sentiments and richer opinions. This difference is attributed to the widespread use of Facebook in Taiwan (see Additional file 1, Table S1) and the easy-to-use characteristic and emotional unambiguity of the "Like" function. Another interesting finding is that the most influential events during the Taiwan election campaign are all closely related to cross-strait relations (see *Influential Events* in *Results*). In particular, in line with the findings in [44], the events more closely associated with public sentiment (such as the Chou Tzu-yu flag incident) appear to have a greater impact than those with merely political meaning (such as the Xi-Ma Meeting).

We provide accurate prediction and automatic causal analysis of the 2016 Taiwan general election, which illustrates the feasibility of applying a data-driven paradigm for political vote analytics. Although our focus is on Taiwan, the proposed signal fusion approach and the event detection model can be applied to other elections or referendums, especially those using majority rule. Considering the different Internet applications used across countries and areas, we may need to adjust the input online information sources and design new measurements for the new signals. Furthermore, we should consider how the election systems of particular countries or areas differ and require adjustment of the prediction model. For example, the US election system is not a direct election but relies on the Electoral College system with 538 electoral votes. Hence, we have to incorporate information about the states and locations of online users into the prediction. However, this information is often unavailable. Nevertheless, we can still consider online users as the voters for a "virtual" direct election and obtain the predictive results as the popular votes for the candidates, which could still indicate the winner if there is a large difference in vote share among candidates. The recent 2016 US Presidential Election demonstrates the power of voices on social media.

Additional material

Additional file 1: Supplementary Information (PDF 2.6 MB)

Acknowledgements
Dr. Guannan Liu was supported by National Natural Science Foundation of China (NSFC)(71701007). Dr. Junjie Wu was supported by National Natural Science Foundation of China (NSFC) (71725002, 71531001, U1636210, 71471009), and Fundamental Research Funds for Central Universities.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the Figshare repository, https://doi.org/10.6084/m9.figshare.6014159.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
ZX, JW and YT conceived the research. ZX and GL conducted the experiments and analyzed the results. All authors wrote and reviewed the manuscript. All authors read and approved the final manuscript.

Author details
[1]Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, School of Economics and Management, Beihang University, Beijing, China. [2]School of Economics and Management, Beihang University, Beijing, China. [3]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China. [4]Foster School of Business, University of Washington, Seattle, USA.

Endnotes

[a] Internet Usage in Taiwan: Summary Report of October 2015 Survey: https://www.twnic.net.tw/doc/twrp/20160108d.pdf.

[b] Taiwan demographics, http://statis.moi.gov.tw/micst/stmain.jsp?sys=100.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 19 April 2018 Accepted: 31 August 2018 Published online: 12 September 2018

References

1. Asur S, Huberman BA (2010) Predicting the future with social media. In: Web intelligence and intelligent agent technology (WI-IAT), 2010 IEEE/WIC/ACM international conference on, vol 1. IEEE Comput. Soc., Los Alamitos, pp 492–499
2. Steinert-Threlkeld ZC, Mocanu D, Vespignani A, Fowler J (2015) Online social networks and offline protest. EPJ Data Sci 4(1):1
3. Bollen J, Mao H, Zeng X (2010) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8
4. Zheludev I, Smith R, Aste T (2014) When can social media lead financial markets? Sci Rep 4(7489):4213
5. Sakaki T, Okazaki M, Matsuo Y (2013) Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans Knowl Data Eng 25(4):919–931

6. Effing R, van Hillegersberg J, Huibers T (2011) Social media and political participation: are Facebook, Twitter and youtube democratizing our political systems? In: International conference on electronic participation. Springer, Berlin, pp 25–35
7. Metaxas PT, Mustafaraj E (2012) Social media and the elections. Science 338(6106):472–473
8. Graham T, Broersma M, Hazelhoff K (2012) Between broadcasting political messages and interacting with voters: the use of Twitter during the 2010 British and Dutch parliamentary election campaigns. Inf Commun Soc 16(5):692–716
9. Enli GS, Skogerbø E (2015) Personalized campaigns in party-centred politics. Twitter and Facebook as arenas for political communication. Inf Commun Soc 16(5):757–774
10. Williams C, Gulati G (2008) What is a social network worth? Facebook and vote share in the 2008 presidential primaries. American Political Science Association
11. DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More tweets, more votes: social media as a quantitative indicator of political behavior. PLoS ONE 8(11):79449
12. MacWilliams MC (2015) Forecasting congressional elections using Facebook data. PS Polit Sci Polit 48(04):579–583
13. Burnap P, Gibson R, Sloan L, Southern R, Williams M (2015) 140 characters to victory?: using Twitter to predict the UK 2015 general election. Elect Stud 41:230–233
14. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. ICWSM 10:178–185
15. Sang ETK, Bos J (2012) Predicting the 2011 Dutch senate election results with Twitter. In: Proceedings of the workshop on semantic analysis in social media. Assoc. Comput. Linguistics, Stroudsburg, pp 53–60
16. Song M, Kim MC, Jeong YK (2014) Analyzing the political landscape of 2012 Korean presidential election in Twitter. IEEE Intell Syst 29(2):18–26
17. Kang M, Zhong H, He J, Rutherford S, Yang F (2013) Using Google trends for influenza surveillance in South China. PLoS ONE 8(1):55205
18. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google trends. Sci Rep 3:1684
19. Curme C, Preis T, Stanley HE, Moat HS (2014) Quantifying the semantics of search behavior before stock market moves. Proc Natl Acad Sci 111(32):11600–11605
20. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with Web search. Proc Natl Acad Sci 107(41):17486–17490
21. Fell D (2005) Party politics in Taiwan: party change and the democratic evolution of Taiwan, 1991–2004. Taylor & Francis, London
22. Kalman RE (1960) A new approach to linear filtering and prediction problems. J Basic Eng 82(1):35–45
23. MacKinlay AC (1997) Event studies in economics and finance. J Econ Lit 35(1):13–39
24. Jackman S (2005) Pooling the polls over an election campaign. Aust J Polit Sci 40(4):499–517
25. Walther D (2015) Picking the winner (s): forecasting elections in multiparty systems. Elect Stud 40:1–13
26. Fisher SD, Ford R, Jennings W, Pickup M, Wlezien C (2016) From polls to votes to seats: forecasting the 2010 British general election. Elect Stud 41(2):244–249
27. Welch G, Bishop G (2001) An Introduction to the Kalman Filter, pp 127–132. University of North Carolina at Chapel Hill
28. Binder J (1998) The event study methodology since 1969. Rev Quant Finance Account 11(2):111–137
29. Zuo Y, Wu J, Zhang H, Wang D, Xu K (2018) Complementary aspect-based opinion mining. IEEE Trans Knowl Data Eng 30(2):249–262
30. Yaqub U, Chun SA, Atluri V, Vaidya J (2017) Analysis of political discourse on Twitter in the context of the 2016 US presidential elections. Gov Inf Q 34(4):613–626
31. You Q, Cao L, Cong Y, Zhang X, Luo J (2015) A multifaceted approach to social multimedia-based prediction of elections. IEEE Trans Multimed 17(12):2271–2280
32. Gayo Avello D, Metaxas PT, Mustafaraj E (2011) Limits of electoral predictions using Twitter. In: Proceedings of the fifth international AAAI conference on weblogs and social media. AAAI Press, Menlo Park
33. Yasseri T, Bright J (2016) Wikipedia traffic data and electoral prediction: towards theoretically informed models. EPJ Data Sci 5(1):1
34. Gopoian JD, Hadjiharalambous S (1994) Late-deciding voters in presidential elections. Polit Behav 16(1):55–78
35. Henderson M, Hillygus DS (2016) Changing the clockthe role of campaigns in the timing of vote decision. Public Opin Q 80(3):027
36. Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489(7415):295–298
37. Kramer AD, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci USA 111(24):8788–8790
38. Bryan CJ, Walton GM, Rogers T, Dweck CS (2011) Motivating voter turnout by invoking the self. Proc Natl Acad Sci USA 108(31):12653–12656
39. Rogers T, Ten BL, Carney DR (2016) Unacquainted callers can predict which citizens will vote over and above citizens' stated self-predictions. Proc Natl Acad Sci 113(23):201525688
40. Wang Z, Solloway T, Shiffrin RM, Busemeyer JR (2014) Context effects produced by question orders reveal quantum nature of human judgments. Proc Natl Acad Sci USA 111(26):9431–9436
41. Rand DG, Pfeiffer T, Dreber A, Sheketoff RW, Wernerfelt NC, Benkler Y (2009) Dynamic remodeling of in-group bias during the 2008 presidential election. Proc Natl Acad Sci USA 106(15):6187–6191
42. Hofman JM, Sharma A, Watts DJ (2017) Prediction and explanation in social systems. Science 355(6324):486–488
43. De Cristofaro E, Friedman A, Jourjon G, Kaafar MA, Shafiq MZ Paying for likes?: understanding Facebook like fraud using honeypots. IMC'14 Proceedings of the 2014 Conference on Internet Measurement Conference
44. Healy AJ, Malhotra N, Mo CH (2010) Irrelevant events affect voters' evaluations of government performance. Proc Natl Acad Sci USA 107(29):12804