Journal of Ambient Intelligence and Humanized Computing (2021) 12:4305–4313
https://doi.org/10.1007/s12652-019-01378-z

ORIGINAL RESEARCH

Leveraging big data for politics: predicting general election of Pakistan using a novel rigged model

Muhammad Awais[1] · Saeed-Ul Hassan[1] · Ali Ahmed[1]

Received: 12 February 2019 / Accepted: 25 June 2019 / Published online: 1 July 2019

Abstract
Big data analytics have shown a tremendous impact on modern politics—among which the election forecasting modeling is notable that utilizes the large scale heterogeneous data sources, such as polls, surveys, and social media popularity to build prediction models by exploiting the power of machine learning and artificial intelligence. In this article, we present a novel machine learning-based election forecasting model that predicted Pakistan's 2018 General Election with the highest accuracy and won a nation-wide competition. To capture the winning probability of individual candidates in a constituency, the model taped an array of statistics from different data sources. Past election data was employed to mine demographic trends of each party across the districts, Twitter, and approval polls were exploited to snap current popularity levels. By employing Bayesian optimization, the model combined the probabilities from different sources by 'rigging' the results for ten seats as a win, where competition was expected to be one-sided. In contrast to the existing models that only predict the aggregate share of votes for different political parties at the national level, our model also effectively predicted the winning candidates for every national assembly seat. The seat share of political parties in the national assembly was predicted with 83% accuracy. Of the total 270 constituencies, 230 winners were among the top two candidates, predicted by the proposed technique. Our model produces the most accurate results of the election compared to all the opinion polls and surveys held before the election 2018 in the country. We showed that big data tools and techniques coupled with the right mixture of machine learning and artificial intelligence models could have a significant impact on modern day political landscape.

Keywords Big data analytics · ICT · Politics · Election prediction · Machine learning · Bayesian optimization

1 Introduction

Advances in information and communications technologies (ICTs) coupled with the big data ools and technologies have enabled scientific communities to mine online social interactions in more sophisticated ways than ever before (Lytras et al. 2017). Recently, with the proliferation of social media platforms, it has been feasible to process large data sets to mine modern human behavior better (Batista-Navarro et al. 2013). The rise of these new interesting data sets have also brought the computing and cognitive science together for the development of new computational platforms, infrastructures, systems and algorithms (Lytras et al. 2018; Shardlow et al. 2018)—in order to better understand about the human interactions in modern daily life, including politics and variety of stakeholders that are involved in social interactions domestically and beyond state borders.

This article leverages heterogeneous big data sources, including the data from social media plate-forms to build an election forecasting model. Keeping in view that election forecasting is an integral part of modern democracies since these forecasting models provide crucial information about the possibility of a regime change. Several different methods such as surveys, mathematical models, economic indicator, citizen forecasting, etc. are used for this purpose. Some of the recent work (Lewis-Beck and Tien 2012; Dassonneville et al. 2017; Prokop 2018; Lewis-Beck and Tien 2018) employing machine learning (Jahangir et al. 2017), and statistical models for election prediction in the developed world has gained some attention.

Code and supplementary material are available at https://awaisrauf.github.io/election_prediction.

Saeed-Ul Hassan
saeedulhassan@gmail.com

[1] Information Technology University, Lahore, Pakistan

In Pakistan, however, election forecasting is mostly limited to opinion surveys. To change this trend, Ignite (ignite.org.pk: a public sector agency for funding scientific research in Pakistan) along with Red Buffer (redbuffer.net), Deep Links (deeplinks.pk) and Code for Pakistan (codeforpakistan.org) arranged a prediction contest (ProPakistani 2018) for Pakistan's general election, 2018. Challenge of this contest was to predict results of national assembly seats before the election.

Our model utilized data from different sources such as Twitter, local survey polls, historical election votes cast, etc., to yield different winning probability estimates of every candidate on a national assembly seat. The final winning probability of a candidate is then obtained by combining all such estimates, and the resulting hyper-parameters were learned on a labeled training set of national assembly seats. This training set was obtained by pre-classifying the winning candidate on the seats, where the election was expected to be one-sided based on past results and overwhelming on-ground popularity indicated by several surveys. We, therefore, name it the 'Rigged Model.' To learn the hyper-parameters on this set, we used Bayesian optimization. This model won the first position among 85 teams and 450 participants (ProPakistani 2018).

The proposed model was 83% accurate for the prediction of seat share of major political parties and correctly predicted the dominance of the surprise third party. On 230 out of 270 seats, actual winners were among the top two candidates predicted by this model. Moreover, it accurately predicted the winning candidate on 160 seats. This model was also better than all of the public polls conducted before the election. On the flip side, the proposed model was not able to identify the emergence of a new party in a relatively smaller province, which has only 5% share of total national assembly seats.

Elections in Pakistan are held every five years for provincial and national assemblies. The country is a federal parliamentary democratic republic with a Parliament, consisting of a directly elected lower house called National Assembly and an upper house called Senate chosen by National Assembly members. National Assembly also elects Prime Minister of Pakistan who is the in charge of the government machinery. This way, National Assembly elections are most crucial, and hence, the center of attention.

The general election was held in 2018 to elect 270 national assembly members. Each district consists of generally one or more constituencies and each constituency's boundaries are limited to only one district. This year's elections were particularly exciting and challenging to predict due to the wave of the surprise third party in a traditional two-party system. Although 120 political parties participated in the 2018 general election of Pakistan, only five are considered to be significant ones. These major political parties include Pakistan Tehreek-e-Insaf (PTI), Pakistan Muslim League Noon (PML-N), Pakistan People's Party Parliamentarian (PPPP), Muttahida Qoumi Movement(MQM) and Muttahida Majlis Amal (MMA).

Rest of the paper is organized as follows. Section 2 reviews relevant literature, Sect. 3 provides a detail account of the model and data used, Sect. 4 shows results and Sect. 5 gives conclusion and a brief overview of possible future directions.

2 Literature review

Scientific election forecasting started in the early 1980s when researchers tried to predict the US presidential election (Lewis-Beck and Rice 1984). This is primarily employed in the USA and Europe whereas it is also being utilized in the developing nations recently (Dwi Prasetyo and Hauff 2015), (Kagan et al. 2015). Although a variety of models are utilized for election forecasting, we can organize them into two broader categories: (1) statistical models (that take political and economic variables and polls into consideration) and (2) social media popularity based forecasting.

Most of these models forecast the seat share of leading parties or incumbent party on a nationwide scale with a few exceptions, where voter share are predicted at state level (Andreas Graefe 2014).

2.1 A review of statistical model-based election forecasting

In this type of models, seat share is considered a function of several variables e.g., economy, GDP, unemployment rate, popularity, seats in previous elections, polls, etc., and ordinary least square (OLS) is mostly fitted with parameters estimated from previous elections, i.e.,

$$\begin{aligned} p &= f(V_1, V_2, \dots, V_n) \\ &= \alpha_1 V_1 + \alpha_2 V_2 + \dots + \alpha_n V_n. \end{aligned}$$

Here V_i shows i-th variable such as economy, past party position, etc., and α_i shows ith hyper-parameter to weight these variables. Campbell et al. (2017) looks at ten similar models for prediction of 2016 USA presidential election forecasting and Tien and Lewis-Beck (2016) reports that these models were better in forecasting than polls.

These models can be divided into three categories: structural, aggregations and synthetic models (Dassonneville and Lewis-Beck 2014). In structural models, votes are considered as a function of political and economic variables. One famous model is the Standard Political Economy model that has been used for the prediction of the incumbent party's share in elections. This model was opted to predict Dutch election 2017 (Dassonneville et al. 2017) and 2016's US presidential elections (Lewis-Beck and Tien 2016) with impressive performance. Other similar models were also adopted for several other elections (Whiteley 2005; Holbrook 2012).

In aggregation models, votes are defined based on aggregation of recent polls (Blumenthal 2014; Traugott 2014). One example of this model is FiveThirtyEight model (Silver 2018), which successfully predicted the results of several Presidential and Senate elections in the USA. Synthetic models are hybrid of both structural as well as aggregations. This model was used by (Dassonneville and Lewis-Beck 2014) to predict the election of several European states.

2.2 A review of social media-based forecasting

A number of studies utilize data from social media platforms to predict elections. Almost all of these studies leverage Twitter data as only Twitter have an open source API for data scraping. However, the scientific community has raised issues in this process (Gayo-Avello 2012) and have cautioned against the use of Twitter data for a direct inference (Mustafaraj et al. 2011). Many studies show weak correlations between Twitter-based results and actual results (Skoric et al. 2012). This issue intensifies if we consider the epidemic of fake news (Temming 2018) and use of bots on Twitter (Craig Timberg 2018).

Famous methods for the use of Twitter in election prediction are either counting based where counting number of hashtags mentioning a specific party (Feldman 2013; Tumasjan et al. 2010) are considered or user analysis based methods where user data is employed to evaluate their demographic and voting preferences (Mislove et al. 2011).

3 Methodology

Our main objective was to predict the winner of each constituency; therefore, we developed a model that outputs a vector of probabilities of the win for each constituency. This vector shows the likelihood of win for each candidate in a constituency. For instance, if a constituency has five candidates then output of the model might look like: [0.2, 0.32, 0.43, 0.02, 0.03]. Each data source gives one such probability vector for each constituency. We assumed results for certain constituencies based on domain knowledge and employed Bayesian optimization to combine these vectors with having the final result. An overview of the model is presented in Fig. 1. Following subsections explain the data and algorithm of the model.

3.1 Data

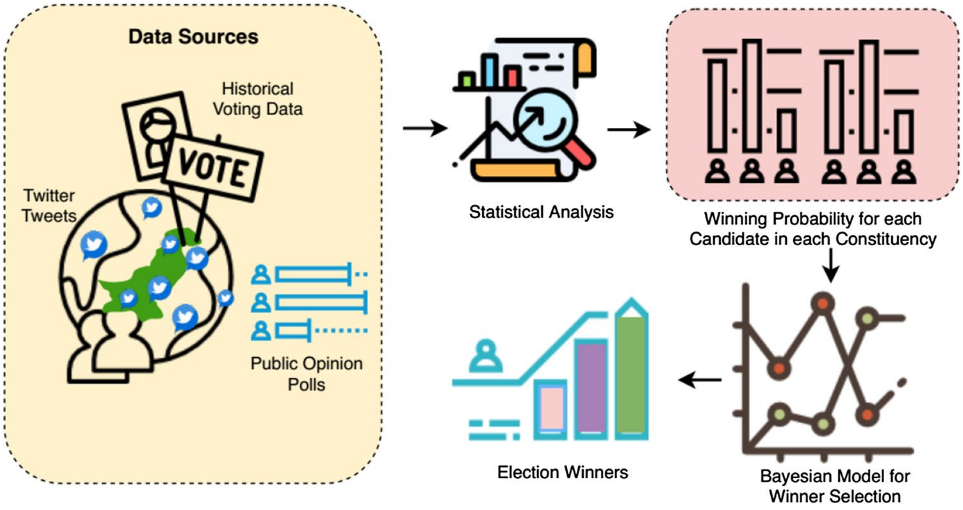
We have leveraged three different types of data in this model: (1) Results of past four elections, (2) public poll data of the last two years and (3) tweets of three weeks before the election.

Past elections data consists of information about each party's vote share in each constituency along with region's information. It is important to note that constituency names and boundaries change in every election, so it is not useful for finding the party's influence in a particular constituency. Therefore, we have converted this data into district level first using regional information and then used it in the model.

Similarly, polls data consists of five different polls conducted in 2017 and 2018 by Gallup Pakistan (Manzar Elahi 2018), (Gallup 2018), Institute for Public Opinion and Research (IPOR 2018) and Dunya TV (Akram 2018).

We also have employed Twitter data collected for four major political parties for three weeks before the election.

Fig. 1 A simplified flow diagram showing generalized version of our proposed model.



For each tweet, we collected its text, the number of times it was retweeted and the number of favorites it got. To collect only relevant tweets, we devised a word profile based strategy along with the location to search Twitter. In this way, we have collected over 640,000 tweets for our analysis. Figure 2 shows the party-wise share of tweets used in our analysis.

3.2 Rigged model for forecasting

Following the tradition of election forecasting models, we considered win probability for a particular candidate as a function of three variables; election history, surveys, and popularity based on social media,

$\mathbf{p_c} = f(\text{election history, surveys, social media}).$

However, contrary to the traditional models, we have predicted results for each constituency, a considerably more challenging problem than finding overall vote share of major political parties. We can formulate our model as follows

$$\mathbf{p_c} = \sum_{j=1}^{J} \alpha[j]h(j,c) + \sum_{k=1}^{K} \beta[k]s(k,c) + \gamma\mathbf{t} + \delta\mathbf{q} \tag{1}$$
$$w_c = \arg\max(\mathbf{p_c}),$$

where $\mathbf{p_c}$: probability vector for c-th constituency, here $\mathbf{p_c} \in \mathbb{R}^{n_c}$, n_c: total number of candidates in c-th constituency, J: total number of past elections used in the model,K: total number of surveys used in the model, $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \delta$: hyper-parameters vectors containing $\alpha[i], \beta[j], \gamma$ as its elements where $0 \le \alpha[i], \beta[j], \gamma, \delta \le 1$ and $\boldsymbol{\alpha} \in \mathbb{R}^J$, $\boldsymbol{\beta} \in \mathbb{R}^K$, $h(j, c)$: function which returns probability vector for a particular constituency c based on one past election j, $s(k, c)$: function which returns probability vector for a particular constituency c based on one poll k, $\mathbf{t}$: probability vector from Twitter data, $\mathbf{q}$: overall likelihood of candidates based on all the previous elections, w_c: wining candidate

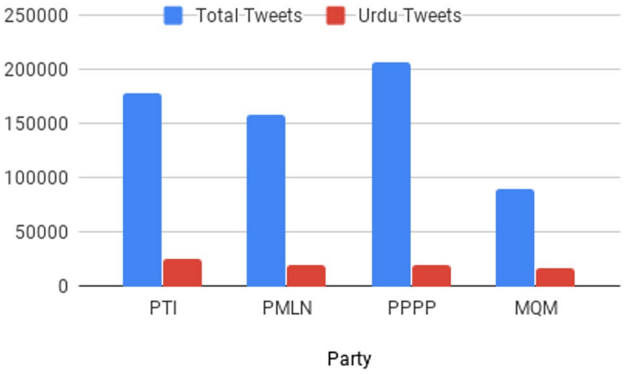


Fig. 2 Number of tweets that were used in the Twitter analysis for four major parties.

In this model, each data source produces a probability vector for each constituency. The process for the computation of this probability vector is explained in next sections. Bayesian optimization is then employed to find optimal values of hyper-parameters such as $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \delta$ to combine these vectors for a final result. Following subsections explains proposed model in detail.

3.2.1 Historical trends and surveys

Results of four elections from 1997 to 2013 were employed to find party influence. This data was utilized to find

1. How likely is a specific party's candidate to win in a district.
2. How likely is a specific party's candidate to win a national assembly seat,

Likelihood of party's candidate in an area is computed through the vote share of that party in previous elections. These statistics were taken at the district level to cope with changing constituency boundaries in each election as explained in Sect. 3.1. For a given constituency c, first its district was found using a function that maps each constituency to its district. Then for a past election j, Equation 2 was used to compute the likelihood

$$h(j,c) = \left[\frac{\sum_{n=1}^{N} v_{m,n,j}}{\sum_{m=1}^{M}\sum_{n=1}^{N} v_{m,n,j}} \right]_{m=1}^{M}, \tag{2}$$

where M: total number of parties participating in election from c-th constituency, N: total number of constituencies in the district of c-th constituency, $v_{m,n,j}$: votes casted for m-th party in n-th constituency for j-th election.

To find overall likelihood of candidates, we have used all the past elections data. Equation 3 was employed to compute overall likelihood.

$$\mathbf{q} = \left[\frac{\sum_{j=1}^{J}\sum_{n=1}^{N} v_{m,n,j}}{\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} v_{m,n,j}} \right]_{m=1}^{M}. \tag{3}$$

Another important data source is surveys that were conducted before the election. Surveys snap popularity levels of major political parties. Each survey gives us province level popularity of major political parties. We have taken these popularity levels as probabilities for each candidate.

The function $s(k, c)$, first finds province of the constituency c and then takes its popularity level from k-th survey.

3.2.2 Social media popularity

To leverage social media in this model, we have collected tweets being tweeted about different parties and employed sentiment analysis to understand underlying support for a party.

Twitter only has an option to search for different keywords and download the tweets related to that search, so we have made word profile for each party consisting of a vector of words that uniquely describes a party. This word profile consists of party name, abbreviations, other common names of the party, major politicians affiliated with the party and province where they had the government in the past term.

Our software developed in python then mined tweets several times a day for three weeks before the election. Each tweet was then analyzed for its sentimental analysis to get the polarity score. Polarity score was between − 1 and 1 where 1 shows extreme positive emotions, − 1 shows extreme negative emotions and 0 shows a neutral response. These scores along with favorites and retweets count were used to find the party's popularity level as shown in Algorithm 1.

```
Result: popularity score of all parties
for party in parties do
    popularity score of party = 0
    for word in party's word profile do
        tweets = download tweets(word);
        if language(tweets) == Urdu then
            tweets = translate(tweets)
        else
            continue
        end
        polarity = sentiment analysis(tweet);
        popularity score = ps (polarity, retweets count, favorite counts)
        popularity score of party += popularity score
    end
end
```

Algorithm 1: Algorithm to collect relevant tweets from Twitter and calculate popularity score.

We used the following equation to find the contribution of each tweet in overall popularity score of a party. The constants with f_c and r_c are based on trial and error.

$$ps = p \times (0.02f_c + 0.01r_c) \tag{4}$$

where ps: popularity score, p: polarity, f_c, r_c: favorite counts given by Twitter API, re-tweet counts given by Twitter API.

A count of re-tweets and favorites for popularity score and emotion or polarity score is used to find how favorable a tweet is for a party. Analysis of a total of 640,000 tweets belonging to four major parties was used in three weeks before the election. Figure 2 shows stats for tweets used in this study.

3.2.3 Bayesian optimization and rigged model

After collecting all the statistics, the next objective of the model is to combine them. One naive approach could be treating all the sources equally. This approach is certainly inefficient as we have prior knowledge that some sources are more credible than others. Another possible approach is to find optimal hyper-parameters for Eq. 1 using historical data as used by Dassonneville et al. (2017). It is not possible to use this approach in the model as we lack similar data for previous elections. Such an approach would also be futile to forecast the effect of the third party which was the main concern in this election. Due to these challenges, we introduced a Bayesian optimization based, novel approach. The success of the model is largely attributed to this approach.

In this approach, we first tracked constituencies where the election was one-sided. For this, we used common knowledge that prominent leaders of major political parties always choose 'safe constituencies'. We rigged the results in the model for these strong candidates and declared the winners. After the rigging, we defined a function $g(\cdot)$ based on Eq. 1 which returns ℓ1-normed difference between predicted and real results for rigged seats given values of hyper-parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \delta)$. We propose finding the values of the hyper-parameters by minimizing the following ℓ_1 objective

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \left[\sum_{c=1}^{l_r} \|p_c(\mathbf{x}) - \mathbf{r_c}\|_1 \right], \tag{5}$$

where $p_c(\mathbf{x})$ is a new function which returns $\mathbf{p_c}$ defined in Eq. 1 given hyper-parameters: $\mathbf{x} = [\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \delta]$, l_r is the number of rigged seats and r_c is one-hot encoded, rigged probability vector with one for rigged winner and zeros for all other candidates.

$$r_c[i] = \begin{cases} 1, & \text{if } i = \text{rigged winner candidate} \\ 0, & \text{otherwise} \end{cases}$$

We have used ℓ_1 norm in the objective function as we want the error to be sparse. The sparse error means most components of error will be zero which is in line with the electoral system as most of the constituencies consists of a large number of candidates, but the real competition is always between top 3 to 4 candidates.

Now our objective is to find values of hyper-parameters for which g minimizes. The objective function g is a black box, and we can not optimize it with conventional optimization techniques as we do not have its expression and derivatives and evaluation is only limited to querying. Since we can query values from this function, one option is to do a grid search. A grid search is not feasible as it would require many evaluations of function for convergence with 0.01 error. Another option is to use random search (Bergstra and

Bengio 2012) but it too is expensive to converge. To avoid these problems, we opted Bayesian optimization to find optimal hyperparameters (Snoek et al. 2012).

Bayesian optimization is a black box optimization technique to optimize an objective function which is difficult to optimize due to lack of any mathematical structure and expensive to evaluate (Brochu et al. 2010). It finds global optima of a function:

$$\underset{\mathbf{x}\in\mathcal{A}}{\text{maximize}} \quad -g(\mathbf{x}),$$

where $g(\mathbf{x})$ is a continuous objective function with unknown structure and expensive to evaluate. Here $\mathbf{x} \in \mathbb{R}^d$ is d-dimensional vector containing hyper-parameters and $\mathcal{A}$ is a search space for $\mathbf{x}$ defined as $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^d : a_i \leq \mathbf{x}[i] \leq b_i\}$. Since $\mathbf{x}$ is bounded in all d dimensions so our search space is a d dimensional hyper-rectangle. Bayesian optimization assumes objective function to be probabilistic and uses Bayes theorem to estimate objective function. From Bayes theorem, we know that posterior $P(D_{1:t}|g)$ is likelihood $P(g|D_{1:t})$ times prior knowledge $P(g)$.

$$P(D_{1:t}|g) \propto P(g|D_{1:t})P(g),$$

Here $D_{1:t} = \{\mathbf{x}_{1:t}, g(\mathbf{x}_{1:t})\}$ is function evaluation for t samples where $\mathbf{x_i}$ shows i-th sample. We assumed Gaussian process prior which is completely described by its mean $m(\mathbf{x})$ and covariance $k(\mathbf{x}, \mathbf{x}')$ functions.

$$g(\mathbf{x}) \sim GP(m(\mathbf{x}); k(\mathbf{x}, \mathbf{x}')).$$

We assume that we have function evaluations for t samples (one random sample at starting point) and we want to sample next best point $t+1$ then we can write our posterior (Brochu et al. 2010),

$$P(g_{t+1}|D_{1:t}, \mathbf{x}_{t+1}) = \mathcal{N}(\mu_t(\mathbf{x}_{t+1}); \sigma_t^2(x_{t+1})) \tag{6}$$

Based on these mean and variance, we define an acquisition function $u(\mathbf{x}|D)$. Then our goal is to find $*argmax_x u(\mathbf{x}|D)$ i.e. find next point where acquisition function is maximized. For this, we have used Upper Confidence Bound (UCB) which is defined as

$$UCB(\mathbf{x}) = \mu_t + \eta\sigma_t. \tag{7}$$

The UCB tries to find the next point where the mean is high which is in line with our problem and also where variance is high, i.e. unexplored regions where we are most uncertain. η is hyper-parameter used to assign importance to each factor. This way, we first assume a mean and covariance, based on this we find a x_t that maximizes 7. From this, we sample our objective function and update our Gaussian process. We repeated this process for 15 points and used hyper-parameters given by it in the model.

4 Results and discussions

Despite the difficulty of predicting results at constituency level and lack of structured data, this model was able to predict result with 83% accuracy for overall seat share of political parties. On 230 out of 270 national assembly seats, the winner was one of the top two predicted by the model. One major success of the model was to predict the government of the surprise, third party with 99% accuracy. Overall results of the model are shown in Table 1.

4.1 Bias for incumbent party

A fascinating insight of the model is bias for PML-N as it predicted 84 seats while PML-N managed to win 64. This bias can be explained based on previous election results. PML-N has ruled Pakistan for the past three decades off and on. It was the dominant party in the 2013 general election by winning 125 seats. Similarly, proposed model underestimated PPPP by 11 seats which again can be explained by the historical decline of the party until the last election.

4.2 Demographic success of model

Another aspect of results is the success of the model for identifying significant areas for the top three political parties as shown in Figs. 3, 4. As it is clear from the map, proposed model was able to identify major political parties for three provinces, i.e. PTI in KPK, PPPP in Sindh and a neck to neck battle between PTI and PML-N in Punjab. On the other hand, model wrongly attributed PPPP and PTI in Baluchistan province.

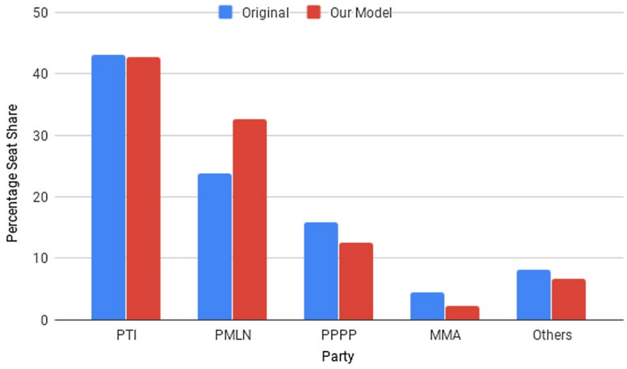
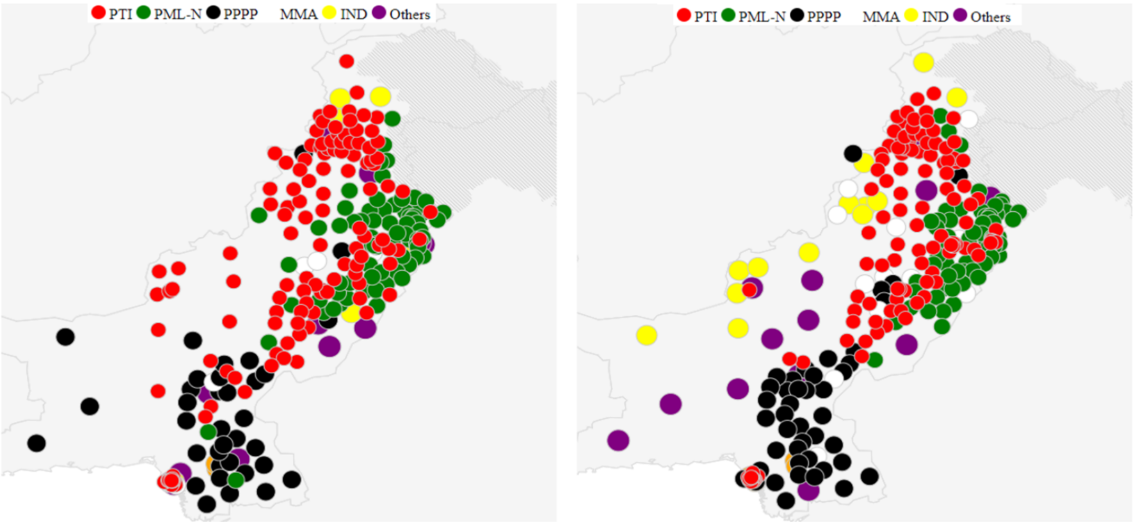


Fig. 3 Proposed model was particularly successful in the prediction of overall seat share of political parties in national assembly

Table 1 Comparison of predicted vs real result

	PTI	PML-N	PPPP	MMA	IND	Others	Total
Predicted	115	88	34	6	11	18	270
Original	116	64	43	12	13	22	270
Error %	0.37	8.88	3.33	2.22	0.74	1.48	17.03


● PTI ● PML-N ● PPPP MMA ● IND ● Others
● PTI ● PML-N ● PPPP MMA ● IND ● Others

(a) Predicted results

(b) Real results

Fig. 4 Comparison of results as dot-map representation on map of Pakistan

4.3 Real time popularity trends from social media

As described earlier, real-time Twitter analysis for popularity trends of major political parties was run for 18 days before the election. Since there was no specified range for popularity value, so we have used relative popularity to show the popularity trends. Similarly, since polarity values in sentiment analysis were between -1 and 1 so we also allowed negative popularity. This increased the utility of the analysis by rightly showing negative sentiment of people on some occasions. For instance, Panama Case(a corruption case against the ruling family) Verdict made popularity of PML-N go down by a lot as shown in the graph. Similarly, another fall in PML-N's popularity can be seen around 15 to 18 July when two significant leaders of the party were arrested on corruption charges (see Fig. 5).

4.4 Comparison with Polls

Polls are a significant medium used for forecasting election. However, polls tend to fail as observed in recent Brexit referendum (Duncan 2018) and Presidential Election of USA (Andrew Mercer 2018). We have observed a similar trend in Pakistan where all the major polls failed to indicate new party's party. The model, on the other hand, was able to predict the seat share of this new party with 99% accuracy. Similarly, this model's cumulative error was way better than both the average error of all the polls and best poll's error. Figure 6 shows a comparison of the model's result with polls (see Table 2).

5 Conclusion

We show that big data analytics may have an enormous influence on modern politics. By exploiting the power of machine learning and artificial intelligence, the election forecasting modeling can better mine heterogeneous data sources, such as polls, surveys, and social media popularity in order to create useful prediction model. In this article, we have discussed our competition winner election prediction model for the 2018 general election of Pakistan. The proposed model was better than all the major surveys conducted before the election. It was able to achieve overall accuracy of 83% for the forecasting of seat share of major parties and predicted third party's government with 99% accuracy. It was also able to predict the winner for 150 out 270 seats while the winner was one of the top two predicted for 230 seats. We also showed the connection between the party's popularity trends on Twitter and real results.

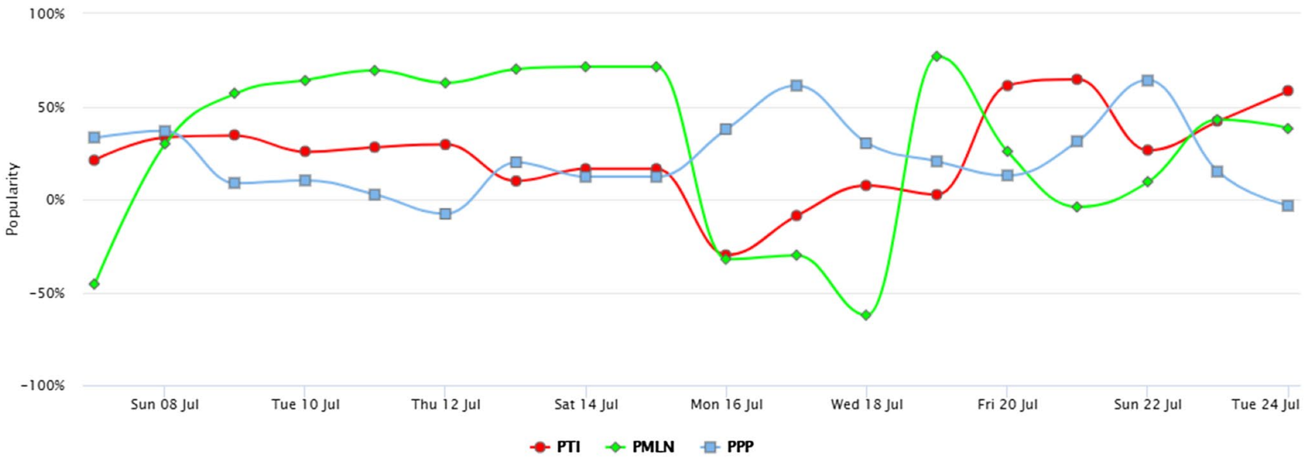


Fig. 5 Relative popularity trends of top 3 political parties of Pakistan by analysis of Twitter from July 7 to July 24, one day prior to general election

Fig. 6 Comparison of error in forecasting of major political party's seat share predicted by different polls vs our model

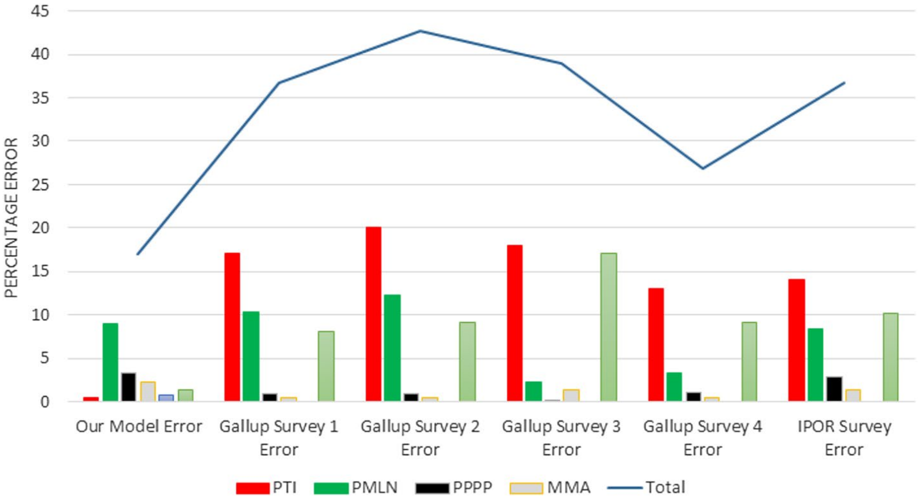


Table 2 Table shows a comparison of major political parties seat share as forecasted by polls and proposed model vs original

	PTI	PML-N	PPPP	MMA	Others
Original	43	23.7	15.9	4.4	8.1
Proposed model	42.6	32.6	12.6	2.2	6.7
Gallulp 2017, 1	26	34	15	4	21
Gallulp 2017, 2	23	36	15	4	22
Gallup 2018, 1	25	26	16	3	30
Gallup 2018, 2	30	27	17	4	22
IPOR 2018	29	32	13	3	23

In the future, we aim to enhance this model by introducing several different improvements such as economic indicators, the influence of electable, use of improved sentiment analysis for Urdu tweets, analysis of media bias, natural language processing techniques (Ananiadou et al. 2013), etc. Last but not least, we show that emerging ICT tools and data generating platforms in combination with appropriate modeling techniques have a significant impact on a modern day political landscape.

Acknowledgements We are very thankful to Muhammad Asim, Iqra Akram and Fahad Shamshad for their insightful discussions and support in data collection.

References

Akram H (2018) Dunya election cell survey 2018 . https://dunyanews.tv/en/Pakistan/449132-Dunya-Election-Cell-Survey-2018-results. Accessed 22 Sep 2018

Ananiadou S, Thompson P, Nawaz R (2013) Enhancing search: events and their discourse context. In: International conference on intelligent text processing and computational linguistics, Springer, pp 318–334

Andreas Graefe AC (2014) State-by-state political economy model. https://pollyvote.com/en/components/econometric-models/jerome-jerome/

Andrew Mercer CDKM (2018) Why 2016 election polls missed their mark . http://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/. Accessed 22 Sep 2018

Batista-Navarro RT, Kontonatsios G, Mihăilă C, Thompson P, Rak R, Nawaz R, Korkontzelos I, Ananiadou S (2013) Facilitating the analysis of discourse phenomena in an interoperable nlp platform. In: international conference on intelligent text processing and computational linguistics, Springer, pp 559–571

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305

Blumenthal M (2014) Polls, forecasts, and aggregators. PS: Polit Sci Polit 47(2):297–300

Brochu E, Cora VM, De Freitas N (2010) A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:10122599

Campbell JE, Norpoth H et al (2017) A recap of the 2016 election forecasts. PS: Polit Sci Polit 50(2):331–338

Craig Timberg ED (2018) Twitter is sweeping out fake accounts like never before, putting user growth at risk. https://goo.gl/meB6pK. Accessed 20 Dec 2018

Dassonneville R, Lewis-Beck MS (2014) Comparative election forecasting. synthetic models for europe. In: Conference on methodological innovations in the study of elections in Europe and beyond, College Station

Dassonneville R, Lewis-Beck MS, Mongrain P (2017) Forecasting dutch elections: an initial model from the March 2017 legislative contests. Res Polit 4(3):2053168017720023

Duncan P (2018) How the pollsters got it wrong on the EU referendum. https://www.theguardian.com/politics/2016/jun/24/how-eu-referendum-pollsters-wrong-opinion-predict-close. Accessed 22 Sep 2018

Dwi Prasetyo N, Hauff C (2015) Twitter-based election prediction in the developing world. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media, ACM, pp 149–158

Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56(4):82–89

Gallup (2018) Elections exclusive: 3 poll results in, Who will you vote for Pakistan?. https://goo.gl/HYiTZX. Accessed 22 Sep 2018

Gayo-Avello D (2012) "I wanted to predict elections with twitter and all i got was this lousy paper"–a balanced survey on election prediction using twitter data. arXiv preprint arXiv:12046441

Holbrook TM (2012) Incumbency, national conditions, and the 2012 presidential election. PS: Polit Sci Polit 45(4):640–643

IPOR (2018) National Survey of Current Political Situation in Pakistan. http://ipor.com.pk/wp-content/uploads/2018/07/National-Survey-of-Current-Political-Situation-in-Pakistan.pdf. Accessed 22 Sep 2018

Jahangir M, Afzal H, Ahmed M, Khurshid K, Nawaz R (2017) An expert system for diabetes prediction using auto tuned multi-layer perceptron. In: 2017 Intelligent Systems Conference (IntelliSys), IEEE, pp 722–728

Kagan V, Stevens A, Subrahmanian V (2015) Using twitter sentiment to forecast the 2013 pakistani election and the 2014 indian election. IEEE Intell Syst 1:2–5

Lewis-Beck MS, Rice TW (1984) Forecasting presidential elections: a comparison of naive models. Polit Behav 6(1):9–21

Lewis-Beck MS, Tien C (2012) Election forecasting for turbulent times. PS: Polit Sci Polit 45(4):625–629

Lewis-Beck MS, Tien C (2016) The political economy model: 2016 us election forecasts. PS: Polit Sci Polit 49(4):661–663

Lewis-Beck MS, Tien CP (2018) House forecasts: structure-x models for 2018. PS: Polit Sci Polit 51(S1):17–20

Lytras M, Aljohani NR, Hussain A, Luo J, Zhang JX (2018) Cognitive computing track chairs' welcome & organization. In: Companion of the The Web Conference 2018 on The Web Conference 2018, International World Wide Web Conferences Steering Committee, pp 247–250

Lytras MD, Raghavan V, Damiani E (2017) Big data and data analytics research: from metaphors to value space for collective wisdom in human decision making and smart machines. Int J Semantic Web Inf Syst (IJSWIS) 13(1):1–10

Manzar Elahi SH (2018) PML-N remains most popular party, Nawaz most favourite leader: survey . https://www.geo.tv/latest/169121. Accessed 22 Sep 2018

Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN (2011) Understanding the demographics of twitter users. ICWSM 11(5th):25

Mustafaraj E, Finn S, Whitlock C, Metaxas PT (2011) Vocal minority versus silent majority: Discovering the opionions of the long tail. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, IEEE, pp 103–110

Prokop A (2018) The terrifying uncertainty at the heart of fivethirtyeights election forecasts. https://www.vox.com/2018/10/24/18009356/fivethirtyeight-nate-silver-election-2018-forecast-analysis. Accessed 20 Dec 2018

ProPakistani (2018) First ever election prediction contest in Pakistan concludes. https://propakistani.pk/2018/08/01/first-ever-election-prediction-contest-in-pakistan-concludes/. Accessed 22 Sep 2018

Shardlow M, Batista-Navarro R, Thompson P, Nawaz R, McNaught J, Ananiadou S (2018) Identification of research hypotheses and new knowledge from scientific literature. BMC Med Inf Decis Making 18(1):46

Silver N (2018) How fivethirtyeights house, senate and governor models work. https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/. Accessed 20 Dec 2018

Skoric M, Poor N, Achananuparp P, Lim E-P, Jiang J (2012) Tweets and votes: A study of the 2011 singapore general election. In: System Science (HICSS), 2012 45th Hawaii International Conference on, IEEE, pp 2583–2591

Snoek J, Larochelle H, Adams R (2012) Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems, pp 2951–2959

Temming M (2018) How twitter bots get people to spread fake news. https://www.sciencenews.org/article/twitter-bots-fake-news-2016-election. Accessed 20 Dec 2018

Tien C, Lewis-Beck MS (2016) In forecasting the 2016 election result, modelers had a good year. pollsters did not. USApp–American Politics and Policy Blog

Traugott MW (2014) Public opinion polls and election forecasting. PS: Polit Sci Polit 47(2):342–344

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. ICWSM 10(1):178–185

Whiteley PF (2005) Forecasting seats from votes in british general elections. Br J Polit Int Rel 7(2):165–173

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