JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A MATHEMATICAL STUDY ON GENERAL ELECTION 2024 OPINION POLLS OF INDIA USING PRINCIPAL COMPONENT ANALYSIS

[1]K Ramakrishnan, [2]V Harish, [3]A Raja Mohammed

[1]Professor of Mathematics,
Department of Science and Humanities,
Karpagam College of Engineering,
Othakkal Mandapam, Coimbatore- 641 032.
Tamilnadu, INDIA.

Abstract: Opinion polls for the 2024 General Election of India have been taken for the study. In this study, the variations between several opinion polls published by different leading agencies have been considered. To identify the most significant opinion polls out of fourteen polls published from October 2023 to April 2024, the method of principal component analysis (PCA) has been employed. The coefficients of correlations between all these polls were established. Loadings (Eigenvalues) of the correlation matrix were found, and the sum of squares of the loadings was tabulated. Finally, the principal components of these opinion polls are calculated. It is observed that the principal components are not equally treated in all the polls. Hence, the Varimax rotation principle has been adopted to ensure equal treatment in opinion polls. The most significant opinion polls are identified, and an opinion poll with the larger variation is also observed. This kind of methodology will help us identify the most significant opinion polls and may be used to identify an overall better prediction model in the future.

Index Terms - Opinion Polls, General Election, India and Principal Component Analysis**.**

I. INTRODUCTION

The Indian general elections are held once every 5 years to elect the Member of Parliament from 543 constituencies all over the country. These elected MPs will elect the Prime Minister, who will rule the country for the next five years. The country had elections in 1952, 1957, 1962, 1967, 1971, 1977, 1980, 1984, 1989, 1991, 1996, 1998, 1999, 2004, 2009, 2014, 2019, and now 2024, respectively. Until the 2009 elections, the election campaigning was party- and ideology-centric. Since 2014, the trend has shifted towards prime minister candidates. There are many processes carried out by various independent agencies to predict the outcome of polls in the past. Pre-poll analysis was conducted weeks or months before the election, and the exit poll was conducted on the day of election, with the voters returning from the voting booth as sample space. This is done in all 7 phases of the election conducted in various parts of the country with different timetables. The widespread use of social media platforms for self-expression, communication, and social participation has resulted in an abundance of voluntarily disclosed personal information online, which can be aggregated to gauge public opinion unobtrusively. As compared to traditional methods of public opinion measurement, social media allows time- and cost-effective data collection and analysis with less human effort. Scholars analyzing social media data to gauge public opinion have supported the idea that the predictive validity of social media analysis does not necessarily rely on how representative the users are of the general population. Before elections, we see a lot of information flying

around: graphs, tables, opinion polls, etc. Numerous literature can be found in research works to understand the sentiment of voters/outcome of the elections results (Tumasjan *et al.* [2011]; Wang *et al.* [2012]; Gayo-Avello [2013]; Razzaq *et al.* [2014]; Miranda Filho *et al.* [2015]; Castro *et al.* [2017]; Pedro Santander *et al.* [2018]; Cury [2019]; Jaidka *et al.* [2019]; Joseph [2019]; Olivares *et al.* [2019]; Oyebode and Orji [2019]; Agarwal and Bansal [2020]; Batra *el al.* [2020]; Belcastro *et al.* [2020]; Brito and Adeodato [2020]; Chan *et al*. [2020]; Grimaldi [2020]; Levi and Patriarca [2020]; Mohbey [2020]; Saxena [2020]; Topîrceanu and Precup [2020]; Bilbao-Jayo and Almeida [2021]; Brito *et al.* [2021]; Chauhan *et al*. [2021]; Reisach [2021]; Stromer-Galley *et al.* [2021]; Sucharitha *et al.* [2021]; Ali *et al*. [2022]; Belcastro *et al*. [2022]; Sancheng Peng *et al.* [2022]). This information tried to predict things like which party will win, who's popular, what issues matter most, etc. But here's the thing: these predictions are just educated guesses, not exact numbers. Sometimes people forget that these percentages and figures are estimates, not concrete facts. For example, if a poll says a party has 40% support with a 5% margin of error, it means the real support could be anywhere from 35% to 45%.

In the present study, an attempt is made to understand which opinion poll from an agency is more significant with the help of principal component analysis.

2. MATERIALS AND METHODS

2.1. Data Collected:

The dataset utilized in this study was compiled from Wikipedia pages dedicated to specific elections, encompassing a diverse range of electoral contexts of India in 2024. Key metrics such as the number of seats allotted to each political party and their corresponding winning percentages were systematically gathered from these pages by the agencies mentioned in Wikipedia. Wikipedia's collaborative nature and extensive coverage of electoral events make it an invaluable resource for researchers seeking comprehensive and up-to-date election data. Moreover, Wikipedia often aggregates information from reputable sources, including official election commissions, news outlets, and academic publications, ensuring the reliability and accuracy of the dataset. In this study, 14 samples of opinion polls published by six leading channels, namely India TV – (CNX) (S1, S4, S11), News 18 (S2), ABP News –(C Voter) (S3, S9), India Today – (C voter) (S7, S14), Times Now (ETG) (S5, S8, S10, S12, S13), and Zee News (Matrize) (S6), at different months' time from October 2023 to April 2024, have been taken.

2.2. Methods:

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. So, the idea is that 5-dimensional data gives you 5 principal components, but PCA tries to put the maximum possible information in the first component, then the maximum remaining information in the second, and so on.

In this technique, one can derive the principal components from the correlation matrix R. This means that the variables have been standardized to have unit variance, and the components turn out to be eigenvectors of R. After computing the variances (eigenvalues, or latent roots) and principal components (eigenvectors) of a correlation matrix, the usual procedure is to look at the first few components, which, hopefully, account for a large proportion of the total variance. Further, the variables are rotated to obtain new variables (principal components or principal axes), and later the number of principal components is reduced by eliminating some relatively unimportant components. Sometimes the first few principal components selected are rotated to achieve a new set of components which can be more easily interpreted. The idea is that each variable should be heavily loaded with as few components as possible. One such technique for accomplishing this transformation is a varimax rotation. This technique tends to eliminate medium-range correlations between the components and the original variables.

3. RESULTS AND DISCUSSION:

The data in Table 1 provides the correlation matrix of the entire fourteen sample opinion poll predictions. Generally, it is seen that all the sample predictions have a strong positive correlation with the remaining, which indicates that the sentiment of the voters is the same in most of the parts of our country. So, to get better information, in the correlation matrix, sample predictions that are highly positively correlated ($i.e., \rho \geq 0.997$) are identified and indicated in bold letters.

.

Table 1 : Correlation matrix of Fourteen opinion polls:

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
S1	1	**0.999**	0.982	0.992	**0.999**	0.988	0.975	0.993	0.950	0.986	0.969	0.964	0.965	0.931
S2	**0.999**	1	0.978	0.994	**0.999**	0.990	0.970	0.995	0.943	0.982	0.963	0.958	0.960	0.924
S3	0.982	0.978	1	0.952	0.983	0.942	**0.999**	0.954	0.991	**0.999**	**0.997**	0.996	**0.997**	0.982
S4	0.992	0.994	0.952	1	0.992	**0.999**	0.941	**0.999**	0.904	0.958	0.931	0.924	0.926	0.880
S5	**0.999**	**0.999**	0.983	0.992	1	0.987	0.976	0.992	0.951	0.986	0.969	0.965	0.966	0.932
S6	0.988	0.990	0.942	0.942	0.987	1	0.930	**0.999**	0.890	0.949	0.919	0.912	0.914	0.864
S7	0.975	0.970	**0.999**	0.941	0.976	0.930	1	0.943	0.995	**0.998**	**0.999**	**0.998**	**0.999**	0.988
S8	0.993	0.995	0.954	**0.999**	0.992	**0.999**	0.943	1	0.907	0.960	0.933	0.926	0.928	0.882
S9	0.950	0.943	0.991	0.904	0.951	0.890	0.995	0.907	1	**0.988**	**0.997**	**0.998**	**0.998**	**0.998**
S10	0.986	0.982	**0.999**	0.958	0.986	0.949	**0.998**	0.960	0.988	1	0.996	0.994	0.995	0.978
S11	0.969	0.963	**0.997**	0.931	0.969	0.919	**0.999**	0.933	**0.997**	0.996	1	**0.999**	**0.999**	0.992
S12	0.964	0.958	0.996	0.924	0.965	0.912	**0.998**	0.926	**0.998**	0.994	**0.999**	1	**0.999**	0.994
S13	0.965	0.960	**0.997**	0.926	0.966	0.914	**0.999**	0.928	**0.998**	0.995	**0.999**	**0.999**	1	0.994
S14	0.931	0.924	0.982	0.880	0.932	0.864	0.988	0.882	**0.998**	0.978	0.992	0.994	0.994	1

The predictions made by the samples, namely S7, S9, S11, and S13, have strong agreement with most of the other predictions. It is observed that there is a significant variation in the predictions of the samples taken in the beginning of 2024 and late 2023. So, it is important to identify the sample predictions that have more variation when compared with others, i.e., principal components. Hence, the method of principal component analysis is employed.

Table. 2 indicates the accumulated contribution rates for various opinion polls done with the help of fourteen samples. It is observed that the prediction made by the sample labeled as S1 has larger loadings (eigenvalues) when compared with other sample predictions. The contribution rate of the prediction made by the sample S2 has a moderate loading. When the percentages of the total variances of the two extracted principal components are accumulated, it may be noticed that these two principal components account for 100% of the variance of the original data. Further, these components were rotated using the varimax rotation method, as shown in Table 3.

To determine at what level the correlation is of importance, a correlation above ± 0.4 is deemed important. These larger correlations are in boldface in the table. It is observed that the first principal component is strongly positively correlated with the sample prediction labeled as S3 and has a moderate positive correlation with S6, S12, and S14. This suggests that these five sample predictions vary together. If one predicts an increase in the number of seats expected to be won by the current ruling party, then the predictions in S6 and S12 tend to increase as well, but the prediction in S14 will decrease. Furthermore, it is seen that the first principal component correlates most strongly with the sample prediction S3, which indicates that this principal component is primarily a measure of S3. It would follow that the communities with high values tend to have a lot of sample predictions in terms of S6, S12, and S14. Whereas communities with small values would have very few of these types of opportunities.

Table 2: Initial Eigenvalues and Extraction of Loadings

Component	Initial Eigenvalues			Extraction Sums of Squares of Loadings		
	Total	% of Variation	Cumulative %	Total	% of Variation	Cumulative %
S1	13.59	97.07	97.07	13.59	97.07	97.07
S2	0.41	2.93	100	0.41	2.93	100
S3	0.00	0.00	100			
S4	0.00	0.00	100			

S5	0.00	0.00	100			
S6	0.00	0.00	100			
S7	0.00	0.00	100			
S8	0.00	0.00	100			
S9	0.00	0.00	100			
S10	0.00	0.00	100			
S11	0.00	0.00	100			
S12	0.00	0.00	100			
S13	0.00	0.00	100			
S14	0.00	0.00	100			

Table 3: Eigenvalues, Contribution rates and Accumulated contribution rates of Principal Components

Agencies	Component Loadings		Rotated Loadings		Communalities
	PC1	PC2	PC1	PC2	
S1	0.0694	-0.0392	**0**	0	0
S2	0.0836	-0.0315	**0**	0	0
S3	-0.0497	-0.0395	**0.9347**	0.2372	0.9299
S4	0.1572	0.0299	-0.2668	0.0463	0.0733
S5	0.0677	-0.0400	-0.1662	-0.0077	0.0276
S6	0.1808	0.0566	**0.5657**	-0.2913	0.4048
S7	-0.0711	-0.0249	0.1108	0.007	0.0123
S8	0.1530	0.0254	0.2918	**0.9425**	0.9734
S9	-0.1234	0.0358	-0.1640	-0.0078	0.026
S10	-0.0376	-0.0455	0.1244	0.0074	0.0155
S11	-0.0873	-0.0102	-0.0304	-0.0022	0.00009
S12	-0.0972	0.0005	**-0.4118**	0.1113	0.1819
S13	-0.0941	-0.0029	-0.1232	-0.0074	0.0152
S14	-0.1511	0.0856	**0.4915**	-0.1244	0.2570

It is also noted that the second principal component increases with the sample prediction S8 for high communities. Hence, from the principal component analysis, it is concluded that the predictions made by the samples S3, S6, S12, S14, and S8 are significant and would be better.

4. CONCLUDING REMARKS:

1. Our study utilized principal component analysis (PCA) to gain insights into election prediction models.
2. Through PCA, we effectively identified underlying patterns and relationships among a multitude of predictor variables.
3. Our analysis revealed the significant role of certain private news agencies as predictors in shaping electoral outcomes. It is observed from the correlation matrix that the prediction agencies labeled S7, S9, S11, and S13 have good agreement with other agencies.
4. Importantly, PCA facilitated the identification of key components driving variation in election data, enabling a clearer understanding of the factors influencing election results. The estimation of loading depicted that the prediction agencies labeled S1 and S2 have the maximum variation with the other remaining agencies, as identified by the principal components.

5. In this analysis, it is seen that the predictions made by agencies S1 and S2 have much variation from others and are to be taken seriously. Hence, the principal component analysis suggested that the qualitative analysis has to be carried out with these two agencies.

5. The predictions made by the samples S3, S6, S8, S12, and S14 have strong or moderate correlations under principal component analysis and significance.

6. By leveraging PCA and similar dimensionality reduction methods, stakeholders can streamline complex election data, enhance predictive accuracy, and ultimately improve decision-making in the political arena.

REFERENCES

[1] Agarwal, A. and Bansal, V. 2020. Exploring sentiments of voters through social media content: A case study of 2017 assembly elections of three states in India. Proceedings of the 22nd international conference on enterprise information systems (ICEIS 2020), 1: 596–602. https://doi.org/10.5220/0009517105960602

[2] Ali, H., Farman, H., Yar, H., Khan, Z., Habib, S. and Ammar, A. 2022. Deep learning-based election results prediction using Twitter activity. Soft Computing, 26(16):7535–7543. https://doi.org/10.1007/S00500-021-06569-5

[3] Batra, PK., Saxena, A. and Goel, C. 2020. Election result prediction using twitter sentiments analysis. Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE. Waknaghat, India https://doi.org/ 10.1109/ PDGC 50313.2020.9315789

[4] Belcastro, L., Cantini, R., Marozzo, F., Talia, D. and Trunfio, P. 2020. Learning political polarization on social media using neural networks. IEEE, 8:47177–47187. https://doi.org/10.1109/ACCESS.2020.2978950

[5] Belcastro, L., Brand,a F., Cantin,i R., Marozzo, F., Talia, D. and Trunfio, P. 2022. Analyzing voter behavior on social media during the 2020 US presidential election campaign. Soc. Netw. Anal. Min. 12(83):1–16. https://doi.org/10.1007/s13278-022-00913-9

[6] Brito, KDS. and Adeodato, PJL. 2020. Predicting Brazilian and U.S. elections with machine learning and social media data. International Joint conference on neural networks (IJCNN). IEEE Access: 1–8. https://doi.org/10.1109/ IJCNN48605. 2020.9207147

[7] Brito, KDS., Filho RLCS. and Adeodato, PJL 2021. A systematic review of predicting elections based on social media data: research challenges and future directions. IEEE Trans Comput. Soc. Syst. 8(4):819–843. https://doi.org/ 10.1109/ TCSS. 2021.3063660

[8] Bilbao-Jayo, A. and Almeida, A. 2021. Improving political discourse analysis on twitter with context analysis. IEEE Access, 9:104846–104863. https://doi.org/10.1109/ACCESS.2021.3099093

[9] Castro, R., Kuffó, L. and Vaca, C. 2017. Predicting Venezuelan states political election results through Twitter, Fourth International Conference on Edemocracy & Egovernment (ICEDEG), IEEE: 148–153.

[10] Chan, E., Krzyzak, A. and Suen, C. Y. 2020. Predicting US elections with social media and neural networks. Pattern Recognition and Artificial Intelligence. ICPRAI 2020. Lecture notes in Computer Science, 12068: 325–335. https://doi.org/10.1007/978-3-030-59830-3_29

[11] Chauhan, P., Sharma, N., Sikka, G. 2021. The emergence of social media data and sentiment analysis in election prediction. J. Ambient Intell Human Comput. 12(2):2601–2627. https://doi.org/10.1007/s12652-020-02423-y

[12] Cury, RM. 2019. Oscillation of tweet sentiments in the election of João Doria Jr. for Mayor. Journal of Big Data 6(42): 1–15. https://doi.org/10.1186/s40537-019-0208-1

[13] Gayo-Avello, D. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. Soc Sci Comput Rev. 31(6):649–679. https://doi.org/10.1177/0894439313493979

[14] Grimaldi, D. Cely, JD. and Arboleda, H. 2020. Inferring the votes in a new political landscape: the case of the 2019 Spanish presidential elections. Journal of Big Data, 7:58-84. https://doi.org/10.1186/s40537-020-00334-5

[15] Joseph, F.J.J. 2019. Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree. Fourth International Conference on Information Technology (InCIT), Bangkok, THAILAND: 50-53.

[16] Jaidka, K., Ahmed, S., Skoric, M. and Hilbert, M. 2019. Predicting elections from social media: a three-country, three-method comparative study. Asian Journal of Communication, 29(3):252–273. https://doi.org/10.1080/01292986.2018.1453849

[17] Levi, E., Patriarca, F. 2020. An exploratory study of populism: the municipality-level predictors of electoral outcomes in Italy. Economia Politica 37(3):833–875.https://doi.org/10.1007/s40888-020-00191-8

[18] Miranda Filho, R., Almeida, JM. and Pappa, GL. 2015. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM) :1254–1261.

[19] Mohbey, KK. 2020. Multi-class approach for user behavior prediction using deep learning framework on twitter election dataset. Journal of Data, Information and Management, 2(1):1–14.

[20] Olivares, G., Cárdenas, JP., Losada, JC. and Borondo, J. 2019. Opinion polarization during a dichotomous electoral process. Complexity, 2019 :1–9. https://doi.org/10.1155/2019/5854037

[21] Oyebode, O. and Orji, R. 2019. Social media and sentiment analysis: the Nigeria presidential election 2019. IEEE 10th annual information technology, electronics and mobile communication conference: 40– 46. https://doi.org/ 10.1109/IEMCON.2019.8936139

[22] Pedro Santander, Claudio Elortegui, Cristian Gonzalez, Hector Allende-cid and Wenceslao Palma. 2018. Forecasting the Chilean electoral year: using twitter to predict the presidential elections of 2017. Social computing and Social media. Technologies and Analytics. Lecture notes in computer science, 10914. Springer, Cham: 298–314. https://doi.org/10.1007/978-3-319-91485-5_23

[23] Razzaq, MA. Qamar, AM. and Bilal, HSM. 2014. Prediction and analysis of Pakistan election 2013 based on sentiment analysis. IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM 2014). IEEE: 700–703

[24] Reisach, U. 2021. The responsibility of social media in times of societal and political manipulation. European Journal of Operational Research, 291(3):906–917. https://doi.org/10.1016/j.ejor.2020.09.020

[25] Sancheng Peng, Lihong Cao, Yongmei Zhou, Zhouhao Ouyang, Aimin Yang, Xinguang Li, Weijia Jia and Shui Yu .2022. A survey on deep learning for textual emotion analysis in social networks. Digital Communications and Networks, 8(5):745–762. https://doi.org/10.1016/j.dcan.2021.10.003

[26] Saxena, A., Kushik, N., Chaurasia, A. and Kaushik, N. 2020. Predicting the outcome of an election results using sentiment analysis of machine learning. International Conference on Innovative Computing and Communications. Springer, Singapore : 503–516.

[27] Stromer-Galley, J., Rossini, P., Hemsley, J., Bolden, SE., McKernan, B. 2021. Political messaging over time: a comparison of US presidential candidate Facebook posts and tweets in 2016 and 2020. Social media+Society, 7(4):1– 13.

[28] Sucharitha, Y., Vijayalata, Y and Prasad, VK. 2021. Predicting election results from twitter using machine learning algorithms. Recent Adv Comput Sci Commun 14(1):246–256. https://doi.org/10.2174/2666255813999200729164142

[29] Topîrceanu, A. and Precup, R-E. 2020. A framework for improving electoral forecasting based on time-aware polling. Social Network Analysis and Mining, 10: 39-51. https://doi.org/10.1007/s13278-020-00646-7

[30] Tumasjan, A., Sprenger, TO., Sandner, PG. and Welpe, IM. 2011. Election forecasts with Twitter: how 140 characters reflect the political landscape. *S*ocial Science Computer Review, *29*(4): 402–418. https://doi.org/10.1177/0894439310386557

[31] Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Proceedings of the ACL 2012 system demonstrations : 115–120.