Email: editor@ijerst.com or editor.ijerst@gmail.com

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

A Sentiment Analysis of Political Tweets for the Purpose of Public Opinion

[1] Krishna Chaitanya, [2] N. Manoj, [3] Shiva Kumar, [4] Sneha, [5] Bhanu Prakash, [6] Dr. Nidamanuru Srinivas Rao, [7] Mrs. Ch. Prasanna,

[1,2,3,4,5]UG Scholar, Dept. of CS, Narsimha Reddy Engineering College, Maisammaguda, Kompally, Secunderabad, India.

[6] Associate Professor, Dept. of CSE, Narsimha Reddy Engineering College, Maisammaguda, Kompally, Secunderabad, India.

[7] Assistant Professor, Dept. of EEE, Narsimha Reddy Engineering College, Maisammaguda, Kompally, Secunderabad, India.

ABSTRACT

For example, forecasting the outcomes of national and municipal elections requires surveying public opinion. This is a typical example of a job in political analysis. But surveys may be costly to do, and if the sample isn't representative of the community, the findings might be skewed. We provide two novel approaches to using the findings of tweet sentiment analysis in survey research. Using only the results of the tweet sentiment analysis, our initial technique outperformed many well-known algorithms. We also present a new hybrid method that estimates election outcomes using public opinion surveys and Twitter. This approach is used to estimate the outcome of the 2023 Greek national election and allows for more accurate, frequent, and cost-effective public opinion estimates. New opportunities for public opinion estimate using social media platforms were opened up by our technique, which showed less variance from the actual election outcomes than traditional public opinion surveys. Terms related to this index include: political tweet analysis, popularity score, sentiment analysis, and Twitter data.

1. INTRODUCTION

Some of the assertions, thoughts, and ideas that make up public opinion are difficult to put a number on. On the other hand, it seems far easier to quantify the popularity of certain political institutions like political parties and individuals. For every i from 1 to n, there exists a mystery popularity score pi for each of the n (political) entities. The political score, which is basically the voting intention, shows the proportion of individuals who would favor entity (political party or candidate) i above all other entities. This is because political voting is a competitive activity. At the outset, we don't know how popular each political entity is (pi, i = 1,..., n). There are a number of methods for estimating it, such as social media data analysis, traditional methods of population sampling and polling, and so on. We aim for the popularity scores pˆ = [pˆ1,..., pˆn]T, which are estimated using any polling technique, to be as near as feasible to the unknown popularity scores p = [p1,..., pn]T. These numbers can only be known on rare occasions, such during an election or voting process. Given a whole population set P and two subsets Pm and Po, we may say that Pm represents social media activists and Po represents poll takers, respectively. The political entities 1,..., n are often referenced in the political texts produced by each member of the set Pm. It is possible to link certain content to a political figure using hashtags on social media. For any political entity i = 1,..., n, text (e.g., a tweet) numbers ai, bi, and ci may be calculated using text sentiment analysis, which sorts the messages into positive, neutral, and negative sentiment categories. Regressing ˆpm from the sentiment dataset S = n (ˆ ai, ˆ bi, ˆ ci), i = 1,..., n o, is the current data analysis challenge. The relative estimations ˆp, ˆpm, ˆpo will vary due to the fact that sets P, Pm, and Po are independent. We can estimate the popularity scores from S using ˆpo and historical measurements of ˆp, without conducting fresh conventional opinion polls, since voting outcomes are too rare compared to traditional polling results.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

Through the years, conventional polling with set Po has consistently achieved minimal polling errors, which is a satisfactory level of accuracy. Choosing the Po sample set appropriately and asking each question by hand has been an expensive ordeal. Consequently, a quicker and less expensive alternative for determining the distribution of popularity scores ˆpm may be offered via social media sentiment research. With the increasing number of individuals taking to social media to voice their opinions on political issues, there is a chance for more accurate and affordable distribution estimates of popularity scores using social media polls. To gauge popular sentiment on voting intentions, we offered two methods in this work for calculating ˆpm. During the last ten years, political campaigns have made extensive use of Twitter and other social media sites. This prompted researchers to look at the possibility of predicting election outcomes and public opinion based on data uploaded online (tweets in the instance of Twitter). With the 2016 US presidential elections, this tendency began on a grand scale and shown very encouraging outcomes [1]. Other two-party (n = 2) political systems that used the same principles also performed admirably [2, 3]. Method [2] was an attempt to refine the popularity measure suggested in [4] for the purpose of forecasting the outcome of the 2017 French presidential election's final round. It is not easy to extend these strategies to multi-party elections (n > 2).

There were a lot of controversial approaches used to bridge the gap between two-party and multi-party elections. In [5], the idea of a sentiment score was put out as the ratio of topic-related positive to negative remarks. Predicting the outcomes of elections involving two or more political parties has made extensive use of this technique. In [6], the real political terrain for the UK general election of 2010 was mapped. Using Twitter data alone to forecast political party popularity is clearly not going to cut it, according to this research. See also [7] for examples of comparable approaches. Because most systems for predicting general elections using Twitter have failed miserably, a hybrid prediction system has been developed [8]. This system employs an election outcome regression model that takes into account several popularity score variables as input. This system was trained using the techniques outlined in [9] on traditional opinion poll results. In social media polling, data sentiment imbalance is a major problem for study. Political commentary on social media is problematic since it often comes from a small number of individuals who are likely prejudiced. It is now much more difficult to obtain accurate data for election prediction, as our study reveals that just 7% of the tweets collected are favorable. So, we come up with an innovative heuristic estimator for election results that relies heavily on unfavorable tweets. In addition, we present a unique approach for regressing the popularity score distribution using prior conventional polls and election outcomes. Our heuristic technique revealed the lowest deviation, compared to other heuristic methods, from the general election outcomes. The most recent public opinion survey conducted by every major Greek polling firm before to the general elections was also surpassed by our hybrid approach, in addition to all current hybrid methods on literature.

2. POLITICAL POPULARITY SCORE ESTIMATION BASED ON TWEET SENTIMENT ANALYSIS

2.1. Heuristic popularity estimation

All subsequent estimations in this study are based on the popularity scores ˆp, which may be obtained from a population in set Pm or from both Pm and Po, without sacrificing generalizability. This is a heuristic for estimating the popularity scores pi, where i ranges from 1 to n, using sentiment labels on political tweets. We begin by automatically classifying all political tweets as favorable, neutral, or negative based on the party they belong to (as determined by the hashtags) using tweet sentiment analysis [10]. For each political entity (party) i = 1,..., n, let's say that there are a total of ai positive, neutral, and negative tweets represented by the n-dimensional vectors a, b, and c. Consider the total of all tweets, both positive and neutral, as vector d = a + b. To create the heuristic popularity score:

$$\hat{p_i}(\mathbf{c}, \mathbf{d}) = \frac{d_i}{d_\tau} \cdot (c_\tau - c_i). \tag{1}$$

The values of dτ and cτ are Pni=1 and ci, respectively. Each party's positive and neutral comment counts are used to determine how the overall negative tweet count (excluding party i) is distributed in ˆ pi(c, d). We update this heuristic estimator and provide the Political Popularity Score Estimator (PPSE) since the distribution of popularity scores should meet the condition Pn i=1 ˆ pi(c, d) = 1.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

$$\hat{p_i}(\mathbf{c}, \mathbf{d}) = \frac{n \cdot d_i \cdot (c_\tau - c_i) + \mathbf{d}^T \mathbf{c}}{n \cdot c_\tau \cdot d_\tau} \quad (2)$$

2.2. Political popularity score regression from tweet sentiment analysis and past opinion polls

2.2.1. Opinion Poll Trends Regressor (OPTR)

Heuristic popularity score projections have shown encouraging outcomes over the years. Unfortunately, as they were derived without considering ground truth or optimization criteria, their prediction accuracy is sub-optimal. Consequently, they missed out on the golden opportunity presented by Machine Learning techniques. Because of this, we can compensate for the decline in precision by drawing on the findings of traditional public opinion polls conducted in the past. Our goal in creating this regression model was to find out how the positive, neutral, and negative counts (ajt, bjt, cjt, j = 1,..., L, t = 1,..., Λ) changed before two successive public opinion surveys. L is the sum of all recorded traditional public opinion polls, and index j is the order in which the surveys were taken. Our main premise is that if there is a major shift in the data measured on social media, the popularity ratings of the entities should also undergo a considerable change. Now we can establish the input for our model by subtracting the average positive, neutral, and negative numbers from a certain time frame Λ prior to two successive opinion polls:

$$\tilde{\mathbf{a}} = \frac{1}{\Lambda}\left(\sum_{t=1}^{\Lambda} a_{kt} - \sum_{t=1}^{\Lambda} a_{lt}\right), \quad (3)$$

$$\tilde{\mathbf{b}} = \frac{1}{\Lambda}\left(\sum_{t=1}^{\Lambda} b_{kt} - \sum_{t=1}^{\Lambda} b_{lt}\right), \quad (4)$$

$$\tilde{\mathbf{c}} = \frac{1}{\Lambda}\left(\sum_{t=1}^{\Lambda} c_{kt} - \sum_{t=1}^{\Lambda} c_{lt}\right). \quad (5)$$

We estimate the change in popularity ratings by applying this mapping to a simple regression model.

$$\hat{\mathbf{r}} = f(\mathbf{x}; \mathbf{w}), \quad (6)$$

where w is the parameter for the model regression function and x = h^˜a||˜ b||˜ci is the size of the 3n-item input vector. Two successive conventional polls are indicated by variables k and l. Mean Squared Error (MSE) loss is used throughout the training method. The regression model (6) may predict the following popularity score estimate ˆp by combining the preceding measurement with ˆr after being trained on the sample data D = {ˆpok − ˆpol, xkl}, k, l = 1,..., L.

$$\hat{\mathbf{p}}_{t+1} = \hat{\mathbf{p}}_t + \hat{\mathbf{r}}_t, \quad (7)$$

The dot (t) denotes the precise instant in time (often days) that is being computed. For data augmentation purposes, variable k may take on the values k = l + 1 or k = l − 1. This is based on the assumption that if the social media statistics were to move in the other direction, the results of the traditional polls would also change in the opposite direction. In order to address the bias that is inherent in traditional public opinion surveys, we will be employing the difference between two successive surveys, denoted as ˆpol, where k is equal to l raised to the power of 1, rather than relying just on their findings, ˆpok. The approach suggested in [8] is comparable to OPTR. Our technique differs greatly from the latter, which uses characteristics derived by heuristic estimators for regression input and uses real opinion poll predictions for output. Because of this inherent bias in opinion surveys, their findings do not account for the interdependence of the various parts of the political system (party). With respect to the suggested approach, fresh estimates will be provided in accordance with the prior ones, as shown in (7). Due to the inherent bias in traditional public opinion surveys, it is necessary to derive starting values from real election results.

2.2.2. Opinion poll grouping to be used in the regression model

There will always be differences in the results of different businesses' traditional public opinion surveys, and these differences might affect the accuracy of the aforementioned political popularity score estimate. For the purpose of the regression model, we will group similar polls taken over the same time period according to how much they deviate from the estimations given by other polling firms (6). Based on a survey performed by firm z = 1,...,m on date tk, let's say that uzi(tk) is the estimated popularity score of party i. Since polling organizations may not always use the same dates for their surveys, we use the formula to perform linear interpolation for each political entity between two

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

successive surveys done by the same business for a given date t, where tk < t ≤ tk+1.

$$u_{zi}(t) = u_{zi}(t_k) + (t - t_k)\frac{u_{zi}(t_{k+1}) - u_{zi}(t_k)}{t_{k+1} - t_k}. \qquad (8)$$

Next, we may get the total of the Mean Absolute Errors (MAE) eŎ(t) associated with the comparison of the firm Ŏ's polls with those of the other polling businesses on date t:

$$e_\zeta(t) = \sum_{z=1, z\neq\zeta}^{m} \frac{\sum_{i=1}^{n} |u_{\zeta i}(t) - u_{zi}(t)|}{n}, \qquad (9)$$

when z ranges from 1 to m. Another issue that is affecting our regressor is the fact that various firms have varying public opinion estimates for the same time period. The goal was to combine public opinion surveys that were conducted within d days of each other by calculating their weighted average error. Assuming you have an opinion polling issue definition, you may adjust the hyperparameter d accordingly.

3.EVALUATION POLITICAL POPULARITY SCORE ESTIMATION METHODS

3.1. Data Gathering

Between June 25, 2022, and June 25, 2023, a total of 1,001,836 tweets pertaining to six Greek political legislative parties were collected via the use of the Twitter API. Using the Transformer technique, which was suggested in [11], all tweets have been categorized as neutral, positive, or negative. This method has a 79% accuracy rate in sentiment classification when evaluated on ground truth Greek political tweets [12]. Our 35 completed public opinion surveys allowed us to train our regression model (6) and validate our suggested methods throughout the data collection period. The general elections in Greece were conducted on May 21, 2023, and June 25, 2023, as per the Greek constitution.

3.2. Comparison of heuristic political popularity score estimators

In order to test and compare our estimator (PPSE) on political Greek Twitter data with five other heuristic estimators and [2, 4, 5, 8, 13] that were either suggested as estimators or utilized as features in regression models. Therefore, we determined the popularity score for each Greek parliamentary party throughout the data collection period based on the aforementioned estimators. Next, we computed the Mean Absolute Error (MAE) to evaluate the various heuristic estimators. This measure is derived from the average deviation of each estimator and the ground truth, which in this case is the results of the general election. Since the sum of some estimators is not 1 for all n entities, we began by normalizing them:

$$\acute{p_i} = \frac{\hat{p_i}}{\sum_{i=1}^{n} \hat{p_i}}.$$

Table 1. MAE between heuristic estimators' results and the results of the Greek general elections of May(21/5/2023) and June (25/6/2023).

Estimators	300 days		200 days		100 days	
	May	June	May	June	May	June
[5]	21.2%	20.71%	21.23%	20.71%	20.94%	20.1%
[8]	19.17%	18.66%	19.18%	18.38%	18.88%	18.37%
[13]	9.2%	9.94%	9.2%	10.97%	10.47%	11.33%
[4]	10.43%	11.27%	10.43%	11.81%	11.55%	11.98%
[2]	9.35%	8.79%	9.35%	8.2%	9.17%	**7.75%**
PPSE (proposed)	**7.06%**	**7.31%**	**7.15%**	**7.62%**	**7.12%**	7.76%

Table 1 presents the testing results during 3 periods, starting from 21 May 2023 and 25 June 2023 for a time window

Table 2. MAE between estimators (OPTR and method [8]), the last recorded opinion poll from different polling companies and the Greek general elections results 25/6/2023.

Methods/Polling Companies	MAE
METRON ANALYSIS	2.17%
MRB	1.89%
MARC	1.63%
GPO	1.57%
PULSE	1.54%
Method [8]	1.42%
OPTR	**1.09%**

for the two separate election dates, which are 100 to 300 days in the past. When compared to competing estimators, ours performs better on the majority of the examined windows. Notably, our heuristic estimator was the only one to accurately forecast the real party ranking (ND > SYRIZA > KINAL > KKE > ELLINIKI LISI > MERA25) based on the election results. The problem is that no estimate has been able to reliably forecast the actual vote shares. Because

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

surveys are able to choose a more representative cross-section of society than Twitter, this might happen. Because of its superiority in predicting election results, the suggested OPTR technique is, in all practicality, the best option (as will be discussed later).

3.3. OPTR model evaluation

We were unable to acquire the poll results or tweets from the last election since data gathering began in June 2022. Fortunately, we were able to test our strategy since there were two general elections. Our strategy essentially tracked the changes in popularity scores from the first election on May 21, 2023, to the second election on June 25, 2023, using the baseline numbers from that date. We contrasted our procedure with that of other polling firms and the outcomes of other methods [8]. Our suggested technique, the [8] method, and the latest recorded opinion poll of each firm before the election date MAE are shown in Table 2, based on the actual election results of 25/6/2023. It is clear that OPTR is superior to both the method suggested in [8] and all of the traditional opinion surveys conducted in the last two weeks leading up to the election. The findings showed that OPTR was biased since it was trained using noisy opinion surveys conducted before the first election day (21/5/2023). Without using any polls released after 21/5/2023, our approach outperforms all others when combined with the first election's outcome and the noisy samples from which it learnt the political trends. Based on the calculations made from equation (9), Figure 1 displays the error that each company has incurred up until the first election date. This inaccuracy is in line with Table 2, providing further evidence that the variance across polling firms may be used to filter out irrelevant differences in training data.

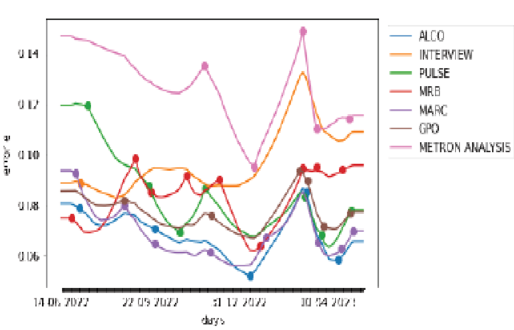

Fig. 1. Poll error (e), from other public opinion polls, throughout the number of days since the 14^{th} of June. Dots indicate the last day when each poll was conducted.

4. CONCLUSION

For the purpose of calculating political popularity ratings using sentiment analysis in social media data, this study presents two novel approaches: a heuristic and a regression technique. Both of them give fairly accurate estimates of how popular politicians are. Although it requires knowledge of previous election outcomes and opinion poll data, the regression-based approach outperforms the heuristic one in terms of accuracy. There is still a significant gap between heuristic popularity estimators and hybrid ones that use social media and previous conventional polls. However, with the advancement of Natural Language Processing (NLP) tools, the accuracy of social media political forecasting should increase. In the meanwhile, hybrid regression techniques are more effective. Nevertheless, our findings show that a hybrid approach combining Twitter data with opinion surveys outperformed traditional opinion polling organizations. Using social media as a main data source, this article suggests a new way to analyze politics. Our study's findings show that this method is a huge step forward in the area, surpassing even the most conventional public opinion surveys.

5. REFERENCES

[1] Brandon Joyce and Jing Deng, "Sentiment analysis of tweets for the 2016 us presidential election," in 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), 2017, pp. 1–4.

[2] Lei Wang and John Q. Gan, "Prediction of the 2017 french election based on twitter data analysis," in 2017 9th Computer Science and Electronic Engineering (CEEC), 2017, pp. 89–93.

[3] Alexandre Bovet, Flaviano Morone, and Hernan A. Makse, "Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump," Scientific Reports, vol. 8, no. 1, pp. 8673, June 2018.

[4] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," Proceedings of the International AAAI Conference on Web and Social Media, 2010.

[5] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith, "From tweets to polls: Linking text

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

sentiment to public opinion time series," International AAAI Conference on Weblogs and Social Media, vol. 11, 01 2010.

[6] Yulan He, Hassan Saif, Zhongyu Wei, and Kam-Fai Wong, "Quantising opinions for political tweets analysis," in LREC, 2012.

[7] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams, "140 characters to victory?: Using twitter to predict the uk 2015 general election," Electoral Studies, vol. 41, pp. 230–233, 2016.

[8] Adam Bermingham and Alan Smeaton, "On using Twitter to monitor political sentiment and predict election results," in Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), Chiang Mai, Thailand, Nov. 2011, pp. 2–10.

[9] Sitaram Asur and Bernardo A. Huberman, "Predicting the future with social media," in 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, vol. 1, pp. 492–499.

[10] Dionisis Karamouzas, Ioannis Mademlis, and Ioannis Pitas, "Public opinion monitoring through collective semantic analysis of tweets," Social Network Analysis and Mining, 07 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[12] Emmanouil Patsiouras, Ioanna Koroni, Ioannis Mademlis, and Ioannis Pitas, "Greekpolitics: Sentiment analysis on greek politically charged tweets," 31st European Signal Processing Conference (EUSIPCO), 09 2023.

[13] Barkha Bansal and Sangeet Srivastava, "On predicting elections with hybrid topic based sentiment analysis of tweets," Procedia Computer Science, vol. 135, pp. 346– 353, 2018.