IEEE Copyright notice

POLITICAL TWEET SENTIMENT ANALYSIS FOR PUBLIC OPINION POLLING

Anestis Kaimakamidis and Ioannis Pitas

Aristotle University of Thessaloniki

ABSTRACT

Public opinion measurement through polling is a classical political analysis task, e.g. for predicting national and local election results. However, polls are expensive to run and their results may be biased primarily due to improper population sampling. In this paper, we propose two innovative methods for employing tweet sentiment analysis' results for public opinion polling. Our first method utilizes merely the tweet sentiment analysis' results outperforming a plethora of well-recognised methods. In addition, we introduce a novel hybrid way to estimate electorally results from both public opinion polls and tweets. This method enables more accurate, frequent and inexpensive public opinion estimation and used for estimating the result of the 2023 Greek national election. Our method managed to achieve lower deviation than the conventional public opinion polls from the actual election's results, introducing new possibilities for public opinion estimation using social media platforms.

Index Terms— Twitter data analysis, Sentiment analysis, Popularity score, Election prediction, Political tweet analysis

1. INTRODUCTION

1.1. Public opinion polling

Public opinion consists of concepts, ideas and statements that seem too abstract to quantify. However, the popularity of certain political *entities* such as, political parties and politicians seem much more easily quantifiable. Let us consider the cause of n (political) entities, each having an unknown popularity score $p_i, i = 1, ..., n$. As political voting is a competitive procedure, the political score (essentially the voting intention) represents the percentage of people that would prefer entity (political party or candidate) i from all other entities. The popularity scores of each political entity: $p_i, i = 1, ..., n$ is initially unknown. It can be estimated in various ways, e.g. through conventional population sampling and polling (through questioning) or by social media data analysis. Any polling method leads to the popularity scores estimation $\hat{\mathbf{p}} = [\hat{p_1}, ..., \hat{p_n}]^T$ that should be as close as possible to the unknown popularity scores $\mathbf{p} = [p_1, ..., p_n]^T$, which can become known only infrequently in special occasions, e.g. through an election/voting procedure. Let $\mathcal{P}$ be the total population set and $\mathcal{P}_m, \mathcal{P}_o$ be the population subsets of a) people that are politically active in social media and b) people participating in a public opinion poll. Each member of the set $\mathcal{P}_m$ produces political texts that, in many cases, refer to the political entities $1, ..., n$ in question. Social media hashtags can be can be used for establishing an association of text to a political entity. Text sentiment analysis can be used to classify such texts into sentiment classes 'positive', 'neutral', 'negative' and quantify their respective text (e.g. tweet) numbers a_i, b_i, c_i respectively for each political entity $i = 1, ..., n$. The data analysis problem at hand is to regress $\hat{\mathbf{p}}_m$ from the sentiment data set $\mathcal{S} = \left\{(\hat{a_i}, \hat{b_i}, \hat{c_i}), i = 1, ..., n\right\}$. As sets $\mathcal{P}, \mathcal{P}_m, \mathcal{P}_o$ differ the respective estimates $\hat{\mathbf{p}}, \hat{\mathbf{p}}_m, \hat{\mathbf{p}}_o$ will be different. Since voting results are too infrequent and transitional polling results are more frequent, we can use $\hat{\mathbf{p}}_o$ and past measurements of $\hat{\mathbf{p}}$ to estimate the popularity scores from $\mathcal{S}$, without performing new traditional opinion polls.

Conventional polling using set $\mathcal{P}_o$ has been satisfyingly accurate over the years, achieving rather low polling errors. However, the process of choosing the sample set $\mathcal{P}_o$ correctly and manually asking political questions is proved to be a costly procedure. To this end, social media sentiment analysis can provide a cheaper and faster alternative solution for estimating the popularity score distribution $\hat{\mathbf{p}}_m$. As more and more people use social media to publicly state their preferences on political topics social media polling provides an opportunity for cheaper and real-time popularity score distribution estimations. In this paper, we proposed a solution for estimating $\hat{\mathbf{p}}_m$ to be used for election result prediction.

1.2. Related work

In the last decade, Twitter and other social media platforms has been widely used as a political communication platform. This urged the scientific community to investigate the idea of generating public opinion and the election results prediction using merely the data posted online (tweets in the case of Twitter). This trend started in a large scale with the US presidential elections of 2016 and showed very promising results [1]. The same methods were implemented for other two-party ($n = 2$) political systems performing equally well [2, 3]. Method [2] tried to improve the popularity metric proposed in [4] for predicting the results of the 2017 French presidential election final round. However, the extension of these methods for multi-party ($n > 2$) elections is not straightforward.

Many approaches were used to bridge the gap between two-party and multi-party elections which resulted in controversial results. *Sentiment score* was proposed as the ratio of positive and negative messages on a topic in [5] along with smoothing the results by using a moving average filter. This method has been widely used for two or multi-party election results' prediction. The mapping of the actual political landscape for 2010 UK general election has been studied in [6]. This study has explicitly concluded that political party popularity cannot be predicted using solely Twitter data. Similar methods have been implemented in [7]. The fact that most methods are trying to predict general election through Twitter produced poor results, led to a hybrid prediction system [8], implementing an election result regression model, whose input comprises several popularity score metrics. This system was trained on conventional opinion poll results by applying the methods described in [9].

A serious research issue in social media polling is data sentiment imbalance. The problem with political comments in social media is that only a few people, that are probably biased, comment positively about a party. Our analysis shows that only 7% of tweets gathered are positive, thus the difficulty of extracting correct election predictions results is significantly increased. Therefore, we propose a novel election result heuristic estimator based primarily on negative tweets. In addition, we propose a novel method for regressing the popularity score distribution output using past traditional polls and election results.

2. POLITICAL POPULARITY SCORE ESTIMATION BASED ON TWEET SENTIMENT ANALYSIS

2.1. Heuristic popularity estimation

Without loss of generality, the rest of this paper estimates refer to the popularity scores $\hat{\mathbf{p}}$, either using population from set $\mathcal{P}_m$ or from $\mathcal{P}_m$ and $\mathcal{P}_o$. The popularity scores $p_i, i = 1, ..., n$ can be heuristically estimated from sentiment labeled political tweets as follows. Firstly, we perform tweet sentiment analysis [10] and automatically tag each political tweet corresponding to a political party (as identified by the tweet hashtags) as positive, neutral or negative.

Let, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ be n - dimension vectors where a_i, b_i, c_i represent the total number of positive, neutral, negative tweets for a political entity (party) $i = 1, ..., n$. Let vector $\mathbf{d} = \mathbf{a} + \mathbf{b}$ represent the sum of positive and neutral tweets numbers. The heuristic popularity score:

$$\hat{p_i}(\mathbf{c}, \mathbf{d}) = \frac{d_i}{d_t} \cdot (c_t - c_i) \tag{1}$$

Where, $d_t = \sum_{i=1}^{n}$ and $d_i, c_t = \sum_{i=1}^{n} c_i$. Essentially, $\hat{p_i}(\mathbf{c}, \mathbf{d})$ distributes the total negative tweet count (without the ones of party i) according to each party's own positive and neutral comment numbers. As the popularity score distribution should satisfy $\sum_{i=1}^{n} \hat{p_i}(\mathbf{c}, \mathbf{d}) = 1$, we modify this heuristic estimator accordingly and introduce the Political Popularity Score Estimator (PPSE):

$$\hat{p_i}(\mathbf{c}, \mathbf{d}) = \frac{n \cdot d_i \cdot (\mathrm{c_t} - c_i) + \mathbf{d}^T \mathbf{c}}{n \cdot c_t \cdot d_t} \tag{2}$$

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

2.2.1. Opinion Poll Trends Regressor (OPTR)

Over the years, heuristic popularity score predictions have shown promising results. However their prediction accuracy is sub-optimal as neither ground truth nor optimisation criteria have been used in their derivation. Therefore, they did not advantage of the recent success of Machine Learning methods. Given this fact, we can handle this accuracy loss by resorting to past conventional public opinion poll results and using them as ground truth data. To this end, we implemented a regression model mapping the change of the aforementioned positive, neutral and negative counts $\mathbf{a}_{ji}, \mathbf{b}_{ji}, \mathbf{c}_{ji}, j = 1...L, i = 1, ..., T$ for a certain time window before two consecutive opinion polls, on the difference of these public opinion polls' results. Index j indicates the chronological order of the opinion poll and L is the total number of recorded conventional public opinion polls. Essentially, our method is based on the observation that a significant change in the data measured on social media should result in a relevant change on the popularity scores of the entities. Let us now define the input of our model using the difference of the average positive, neutral and negative counts of a specified time window T before two consecutive opinion polls:

$$\tilde{\mathbf{a}} = \frac{1}{T}\left(\sum_{i=1}^{T} \mathbf{a}_{ki}^T - \sum_{i=1}^{T} \mathbf{a}_{li}^T\right),$$

$$\tilde{\mathbf{b}} = \frac{1}{T}\left(\sum_{i=1}^{T} \mathbf{b}_{ki}^T - \sum_{i=1}^{T} \mathbf{b}_{li}^T\right),$$

$$\tilde{\mathbf{c}} = \frac{1}{T}\left(\sum_{i=1}^{T} \mathbf{c}_{ki}^T - \sum_{i=1}^{T} \mathbf{c}_{li}^T\right).$$

By using this input on a simple regression model we implement this mapping and estimate the change $\hat{\mathbf{r}}$ in popularity scores:

$$\hat{\mathbf{r}} = \mathbf{w}^T \mathbf{x} + \mathbf{b}, \tag{3}$$

where, $\mathbf{x} = \left[\tilde{\mathbf{a}} || \tilde{\mathbf{b}} || \tilde{\mathbf{c}}\right]$ is the input vector of size $3n$. Variables k and l indicate two consecutive conventional polls and $\mathbf{w}, \mathbf{b}$ are trainable regression parameters. Training of $\mathbf{w}, \mathbf{b}$ is performed using the Mean Squared Error (MSE) loss.

Once the regression model (3) is trained on sample data $\mathcal{D} = \{\hat{\mathbf{p}}_{ok} - \hat{\mathbf{p}}_{ol}, \mathbf{x}_{kl}\}, k, l = 1, ..., L$, with L being the total

number of the recorded conventional opinion polls, it can estimate the next popularity score estimate $\hat{\mathbf{p}}$ by adding $\hat{\mathbf{r}}$ to the previous measurement:

$$\hat{\mathbf{p}}_{t+1} = \hat{\mathbf{p}}_t + \hat{\mathbf{r}}_t, \quad (4)$$

t indicates the specific time spot that is being calculated (usually days). Variable k can be both $k = l + 1$ or $k = l - 1$ for data augmentation reasons, assuming that the opposite change on the social media statistics would bring the exactly opposite change to the conventional polls' results.

To counter the bias introduced by conventional public opinion polls, instead of using their results directly $\hat{\mathbf{p}}_{ok}$, we choose to utilise the difference of two consecutive opinions polls $\hat{\mathbf{p}}_{ok} - \hat{\mathbf{p}}_{ol}, k = l \pm 1$. The hybrid political popularity score regression is similar to that in [8]. However, the latter's regression input , consists of a few heuristic estimators, perceived as features and its output utilises the actual opinion poll estimations, differing significantly from our method. Thus, their results fail to analyse the components of the political system (parties) as dependent entities and also do not filter this bias added from opinion polls. As far as the proposed method is concerned, the estimation will be given according to the previous one, as (4) indicates. This creates the need for initial values that can be collected from actual election results, because of the bias introduced from conventional public opinion polls.

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

The above-mentioned political popularity score estimation can be sensitive to the unavoidable variations observed between various conventional public opinion estimation conducted by different companies. The chosen approach is to group different polls, that were conducted on the same period, according to their deviation from the estimates of other polling companies, to be used in regression model (3). Let us suppose that $u_{ij}(t_k)$ is the estimation of the popularity score of party j in a poll conducted by company $i, ..., m$ at date t_k As the poll dates differ across polling companies,we perform linear interpolation for each political entity between two consecutive polls conducted by the same company for a given date t, where $t_k \leq t \leq t_{k+1}$, by using the formula:

$$u_{ij}(t) = u_{ij}(t_k) + (t - t_k)\frac{u_{ij}(t_{k+1}) - u_{ij}(t_k)}{t_{k+1} - t_k}. \quad (5)$$

Then we can compute the Mean Absolute Error (MAE) $e_i(t)$ for company i on date t between the polls of company i and the rest of the m:

$$e_i(t) = \sum_{k=1, k\neq i}^{m} \frac{\sum_{j=1}^{n} |u_{ij}(t) - u_{kj}(t)|}{n}, \quad (6)$$

where, $i = 1, ..., m$. The variations between public opinion estimations by different companies on the same period is also causing problems to our regressor. To this end, when two or more public opinion polls were held less than d days apart, they are merged using the weighted average according to the errors measured. The hyperparameter d is set depending to the specifications of the opinion polling problem.

3. EVALUATION POLITICAL POPULARITY SCORE ESTIMATION METHODS

3.1. Data Gathering

More than 1,000,000 tweets have been gathered about six Greek political parliamentary parties, using the Twitter API from the 25^{th} June 2022 until the 25^{th} June 2023. All tweets have been labelled as neutral, positive or negative using the Transformer method proposed in [11], that exhibits 79% sentiment recognition accuracy, tested on ground truth Greek political tweets [12]. During the data gathering period we managed to collect 35 public opinion polls, that we utilized for training our regression model (3) and also validating our proposed techniques. The Greek general elections were held on 21/5/2023 and on 25/6/2023, according to the provisions of the Greek constitution.

3.2. Comparison of heuristic political popularity score estimators

In order to evaluate and compare our estimator (PPSE) with five different heuristic estimators and [2, 4, 5, 8, 13], that were either proposed as estimators or used as features in regression models, on political Greek Twitter data. To this end, we calculated the popularity score, according to the aforementioned estimators, for every party inside the Greek parliament during the data collection period. Then, in order to compare the different heuristic estimator outputs we calculated the Mean Absolute Error (MAE), defined as the average error of each predictor between the general election results (used as ground truth) and the estimator output. As some estimators do not sum to 1 for all n entities, we normalised them first: $\acute{p_i} = \frac{\hat{p_i}}{\sum_{i=1}^{n} \hat{p_i}}$.

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).

	300 days		200 days		100 days	
Estimators	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 Jun 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**%

of 100 to 300 days backwards, for the two different election dates. It is clear that our estimator outperforms other estimators during most windows tested. It must be noted here, that according to the election results, our heuristic estimator was the only one to correctly predict the actual party ranking (ND > SYRIZA > KINAL > KKE > ELLINIKI LISI > MERA25). However, all the estimators struggle to predict the actual votes shares. This might occur, because of the advantage polls have over Twitter, of picking a balanced sample of different society groups. Hence, for all practical purposes, it is best to use the proposed (OPTR) method, which provided much superior election result prediction, as analysed in the next section.

3.3. OPTR model evaluation

Since the data collection started on June 2022 and we were unable to access neither previous election's poll results nor the tweets of the respective periods. Thankfully, general elections were held twice, allowing us to test our method. Essentially, our method used the results of the elections of 21/5/2023 as initial values and calculated the popularity score changes until the second elections of 25/6/2023. Our method is compared with different poll companies and also the method [8] results. Table 2 presents the estimations, for our proposed method, [8] method and the last recorded opinion poll of each company before the election date MAE from the actual election results of 25/6/2023. As seen, opinion poll regressor (OPTR) outperforms the technique proposed in [8] and also all the conventional opinion polls held on the last two weeks before the election date. OPTR only trained with the noisy opinion polls before the first election date (21/5/2023), that proved to be biased. Although our technique learned the political trends from those noisy samples, their combination with the first election's result surpasses all other techniques and opinion polls, without using any of the polls published after 21/5/2023. Figure 1 presents the deviation of each company from the others until the first election date as calculated from (6). This deviation agrees with Table 2, which is an additional validation for using the deviation between companies to eliminate unwanted variations on training data.

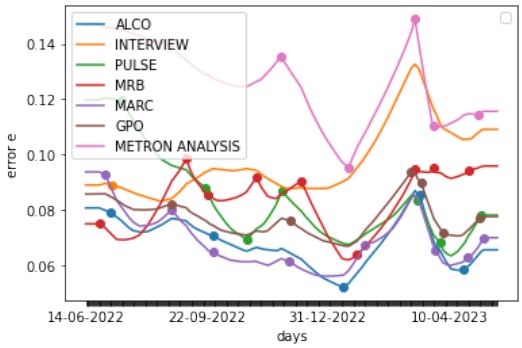

Fig. 1. Poll error (e), from other public opinion polls, throughout the number of days since the 14^{th} of June. Dots indicate the last day when each poll was conducted.

4. CONCLUSION

In this paper, we proposed two new methods for estimating political popularity scores through sentiment analysis of social media data: both a heuristic and a regression method are proposed. They both provide rather good estimation of political popularity scores. The regression based method is more accurate than the heuristic one but requires knowledge of past opinion poll data and past elections results as well. Although, the difference between heuristic popularity estimators and hybrid (using social media and past conventional polls) ones is still considerable, as NLP tools get more advanced the results we get from political forecasting through social media should become more and more accurate, but for the time being hybrid regression techniques outperform them. As indicated by our experiments though, a hybrid method using both Twitter data and opinion polls proved to provide better results than conventional opinion polling companies. This paper introduces a new level on social media political analysis, as it is the first time the public opinion polls have been outperformed by a social media-based technique.

5. ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 951911 (AI4Media). This publication reflects only the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

The authors would like to thank Ioanna Koroni for helping annotate the Greek political tweets ground truth dataset.

6. 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 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.