Political Profiling using Feature Engineering and NLP

Chiranjeevi Mallavarapu[1], Ramya Mandava[1], Sabitri KC[1], Ginger Holt[2]

[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX-75275, USA
[2] Facebook
{cmallavarapu, rmandava, skc}@smu.edu
gingermholt@fb.com

Abstract. Public surveys are predominantly used when forecasting election outcomes. While the approach has had significant successes, the surveys have had their failures as well, especially when it comes to accuracy and reliability. As a result, it becomes challenging for political parties to spend their campaign budgets in a manner that facilitates the growth of a favorable and verifiable public opinion. Consequently, it is critical that a more accurate methodology to predict election outcome is developed. In this paper, it has been identified that dynamic public data from open sources on politicians has influential features that can help in determining the outcome of elections. Our model yielded a 0.71 Pearson Correlation between derived features and percentage of votes predicted. Hence, the election outcomes are also dependent on elements such as a candidate's affiliation with featured organizations, their presence in public events and their online footprint and their associated data in Wikipedia among other factors.

1 Introduction

Outcome of political contests is inarguably one of the topics that receive incredible public attention around the world. News media and political analysts always spend considerable amount of money to find a winning candidate, especially in constituencies where the competition is high. Similar attention is given to election result's forecasts as they are seen by many as a precursor of the actual results. In general, forecast of political outcomes is based on voter polls through multiple mediums such as digital polls, telephone surveys among others. The need for election polls has resulted in the growth of an industry constituted of companies specializing in undertaking public surveys and rank individual candidates based on opinion. Sometimes, these predictions are inaccurate due to voter bias, which has a major impact on campaign planning and budget spending for political parties. By having an unbiased prediction, it will be possible for political candidates and the parties to better manage their campaign expenditure.

In the recent past, political predictions have been inaccurate to the surprise of audience. Media and political analysts identified the problem stemming from the inadequacy of voter opinion and their overall participation at the time of polling[1]. These inconsistencies make the political spending by the parties and candidates ineffective. This paper presents a solution to this problem by collecting dynamic public data from open sources on politicians. There is a significantly large volume of data relevant to this problem available in open access. It has been identified that Googles Knowledge

Graph, Wikipedia, and NewsAPI provides a 360 degree view of data entities that can collectively address this issue. These data sources have been integrated into our analysis through restful APIs.

Feature engineering techniques have been applied and features such as a candidate's affiliation with particular organizations, their participation in public events, their online footprint and their associated data in Wikipedia among other feature considerations showed high power in predicting the percentage of votes the individual politician receives in the election.

Random forest regression model with above mentioned features produced predictions that slabbed at 30 percent. This prompted for a revised model including additional features such as party affiliation and number of opponents for each candidate within their constituency. This second revision eliminated the problem of slabbing at 30 percent. However, the predictions for high performing candidates have high deviation from actual percentage of votes compared to low or medium performing candidates. NLP techniques such as tf and tf-idf have been applied on the Wikipedia data available for these high performing candidates. This third revision of Random Forest model resulted in better overall predictions.

Revision 3 of the model is a Random Forest model performed with 5-fold cross validation and 150 estimators that resulted in a Pearson Correlation of 0.71 and Mean Squared Error(MSE) of 0.019. As this revision 3 is performed only on subset of data i.e. the candidates with Wikipedia page, our final model is an ensemble of revision 3 model for candidates with Wikipedia page, and revision 2 model for candidates without the Wikipedia page. The difference between these two revisions being the feature sets.

Based on our analysis and modeling, it can be concluded that the voting percentage can be predicted based on a candidate's affiliation with featured organizations, their presence in public events and their online footprint and their associated data in Wikipedia mostly having high frequency key words such as 'congress','republican' and 'house'.

The paper is organized into 7 sections. Section 2 provides a brief overview of some of the prior research and notable attempts to predict election results. In Section 3, the data sources and feature identification is explored along with the reason of choosing those particular data sources. Further in this section, he dependencies of features with respect to response have been scrutinized that helped in choosing the right features.Section 4, provides specific details on how each model was optimized and also evaluates them. This section further discusses the revisions of these models. Section 5 provides the results of our models and the analysis of results. This section also identifies the top features having most impact on our predictions. In Section 6, discusses ethical consideration of the project and the associated data. Finally, section 7 presents the conclusion and how our model can be extended to other applications.

2 Prior Research

This section provides a brief overview of some of the notable attempts to predict elections. One of these approaches is sentiment detection, which is not only applicable in the political scene but has also been successfully used in assessing the attitude of a

particular population towards a given product or service. This approach applies a broad range of feature selections on a collection of lingual digital content such as forum postings, reviews, and blogs. Although these source provide some reliable measure of the sentiments, there is a chance of bias. These sentiments can then be mapped to a linear scale that relates to emotional versus neutral, and positive versus negative language.

On the other hand, in the midst of user data privacy concerns[3], NLP also has been extensively applied in the analysis of Twitter feeds as a way of evaluating the sentiment of an audience. In addition to the potential of using features such hashtags, the analysis of sentiment can also be performed by considering other features such as punctuation and n-grams. Elements such as smileys are reliable when it comes to the classification of sentiments. The classification of sentiment can also be performed using different schema. This begins with the annotation of the collected data with consideration of the emoticons expressed in the data, which is followed by the use of distant supervised learning[4].

The Bayes theorem is another approach that shows much potential in the prediction of election results in the US [5]. The approach involves a series of steps. First, a time horizon for analysis is selected and the parameters for the analysis identified. After this, the configuration of the parameters is studied in the context of a typical election year for which the forecasts are prepared followed by the extrapolation of the identified configuration to past years. In the fourth step, the Bayes theorem is applied with the data from previous elections to determine the possibilities of the candidates. Projections can be made for the unknown parameters depending on the results of earlier campaigns.

3 Data Acquisition and exploration

Based on initial assessment of multiple relevant data sources, it has been determined that open schematic data that is constantly updated is the best way to approach this problem. These factors resulted in identifying Google's Knowledge Graph, Wikipedia Data and NewsAPI as the leading data sources in their respective fields.

3.1 Data Collection

Googles Knowledge graph A knowledge graph shows relationships between real world entities and describes them in an organized manner. It has potentially interrelating arbitrary entities with each other across various topical domains [2]. Googles knowledge graph has been leading the industry in this area.

Googles Knowledge graph[5] provides an API which can be used to download certain information about a politician. Google collects public information from multiple sources and uses proprietary algorithms to rank and rate the information and structure it in a standard format that can be pulled using its knowledge graph API. The data is structured as per standard data types defined in standard schema organization[6].Key schema types in relation to politicians have been identified as : Person, Event and Organization.

[3] https://link.springer.com/article/10.1007/s00146-014-0549-4

[4] https://journals.sagepub.com/doi/pdf/10.1177/2053951716645828

[5] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

[6] https://schema.org/

Wikipedia Wikipedia[7] being an authentic open data source, it has been chosen to extract candidate biographic information such as educational background, date of birth, occupation and political history as available on their Wikipedia page. Wikipedia allows for edits from end users from all over the world, hence cross validating the data being entered and makes it authentic.

Wikipedia rest API along with pywikibot is used to get full text of Wikipedia pages as well as standard templates of data from Wikidata for the 2016 Senates across all states. Even though mostly Wikipedias information about a politician is divided into biography, personal section and political career section, Pywikibot gives the necessary tools to extract individual elements. Further, NLP features were extracted from the Wikipedia page of a candidate.

2016 candidate Wikipedia pages Data about the politicians voting outcome has been collected from Wikipedia 2016 senator page, using JSON download to get the total number of votes they received to evaluate the accuracy of our features[3].

Newsapi.org News API[8] has been considered as an additional data source to crawl, index and monitor the top news related to every politician from over 30,000 news sources and blogs. It is a simple HTTP REST API for searching and retrieving any news article from the Internet based on different criteria like keyword, date published, language, domain etc. and can be sorted in different order for e.g.: date published, popularity of source, number of social shares etc. Hence it curates the data from thousands of different sources and serves as a great unbiased data source for our project. Information about the latest news on politician has been gathered from NewsAPI using their REST API.

3.2 Data Exploration

Our consolidated data set includes Wikipedia 2016 senator candidate **names** and **votes**, corresponding Google Knowledge Graph **Events** and **Organizations**, NewsAPI **Articles** and **Wikipedia pages**.

Data exploration has been started with California as our sample state which had 37 candidates that ran for Senate elections in 2016. Out of the 37 candidates, 7 candidates have a Wikipedia page, 13 candidates have presence on Google Knowledge Graph and 17 candidates have been mentioned in articles derived from NewsAPI.Features for further analysis have been defined as **number of organizations** a candidate is associated with, **number of events** a candidate participated in, **number of News articles** associated with a candidate and whether the candidate has **Wikipedia page** and **Google presence**. The data exploration then has been extended to the remaining states and candidates.

A pair-plot based on these derived feature data-set indicated a clear correlation between **number of Votes** and the features identified as shown in Fig.1.

[7] https://en.wikipedia.org/

[8] https://newsapi.org/

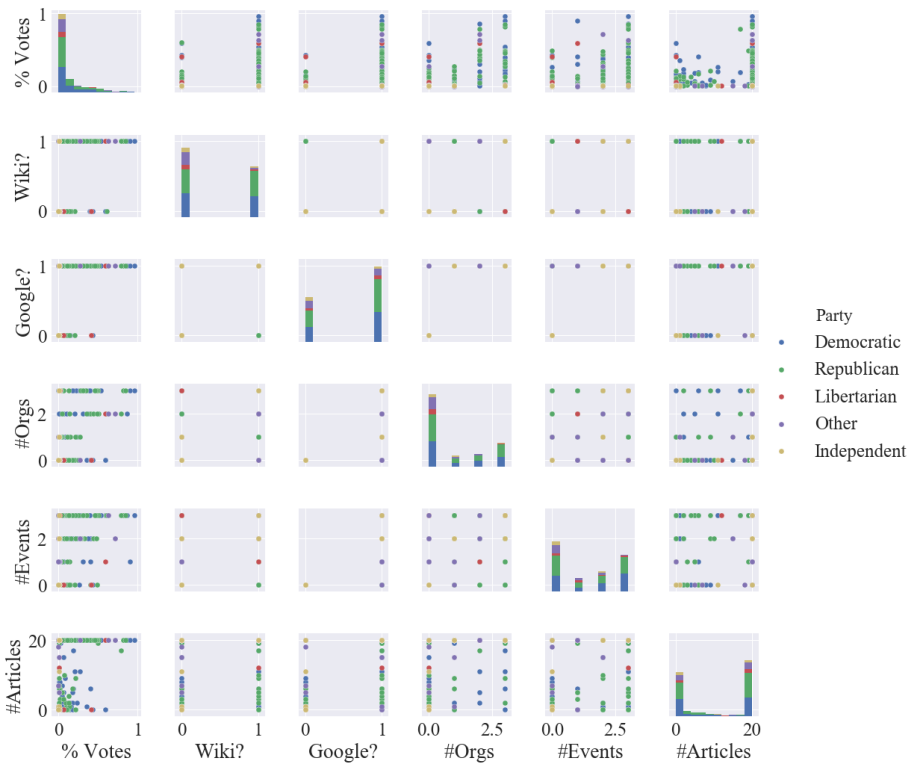

Fig. 1: SNS Pair plot of engineered features

It is apparent that the percentage votes vary based on the number of events and organizations a candidate is associated with. Two different patterns can be seen here on Votes vs Articles: One that doesn't change the votes based on increase in a candidate's no: of news articles, another a positive correlation.

Percentage of Votes vs Orgs As observed in Fig.2, it is noticed that the people that have association with organizations seems to have a clear positive impact on the number of votes. However, it is observed that people that don't have association with organizations also seems to have a minor effect on the votes compared to above scenario. The plot has been color-coded with number of events as color-bar and it can be noted from Fig.2 that the higher the number of organizations, it is also likely for the candidate to have a higher number of events associated with him/her. Few instances can be observed where higher number of orgs are associated with lower number of events for a candidate, hence a need to separately explore events vs percentage of votes.

In addition, from the results, it is possible for a candidate to enjoy a higher percentage of votes with a considerable number of events associated with them and not necessarily the highest number of events and vice versa since several cases of high number of events are associated with lower percentage of votes. It is however clear that, in most

cases, the number of events associated with a candidate contribute to the percentage of votes they get.

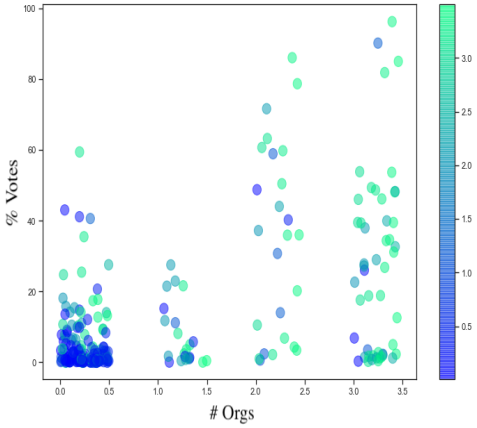

Fig. 2: Scatter plot of number of organization vs. percentage of

The included color scheme based on number of events that candidates are associated with as depicted in Fig.2, helps to understand the cluster of candidates that have non-zero votes with zero association with organizations. It is evident from these plots that, all candidates that have 2 percent or more votes in that cluster have mostly non-zero events, suggesting that events feature is also influential in those cases.

Percentage of Votes vs Events A positive correlation can be observed between the number of events and percentage of votes from Fig.3.

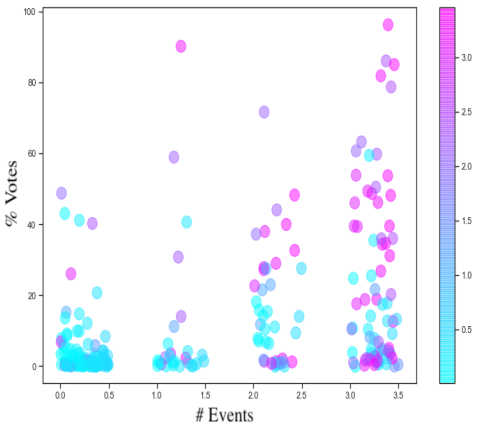

Fig. 3: Scatter plot of No: of Orgs vs. Percentage of Votes - Events

It is also observed that when the number of events is more than 2, candidates associated with 2 or more organizations generally seems to have higher percentage of votes. In addition, from the results, it is possible for a candidate to enjoy a higher percentage of votes with a considerable number of events associated with them and not necessarily the highest number of events and vice versa since several cases of high number of events are associated with lower percentage of votes. It is however clear that, in most cases, the number of events associated with a candidate contribute to the percentage of votes they get.

Percentage of Votes vs Articles Analyzing no: of articles vs. percentage of votes, there is a positive correlation. However, for candidates with no: of articles more than 7.5 and with small percentage of votes as can be seen in Fig.4, it is verified that the articles do not belong to the political candidate in question. The candidates with low percentage of votes and with higher no: of articles seem to have low no: of Events and Organizations associated with them as evident by the color coding in the scatter plot.

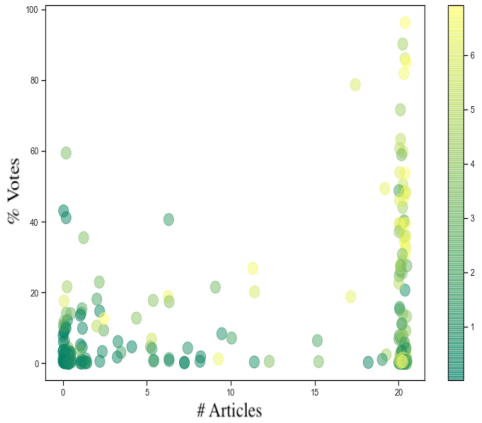

Fig. 4: Scatter plot of No: of Orgs vs Percentage of Votes - Articles

Percentage of votes vs Google presence/Wikipedia presence Analyzing the data for percentage of votes vs. Google presence and percentage of votes vs. Wikipedia presence, there is a general positive correlation.

As it is evident from Fig.5 and Fig.6, candidates having higher number of votes generally tend to have Wikipedia presence. This was a good indication for the possibility of applying NLP techniques for modeling. It is interesting to see that, based on the color coding of the number of Organizations plus number of Events, candidates that have no online presence on Google or Wikipedia also does not have association with noted organizations nor events.

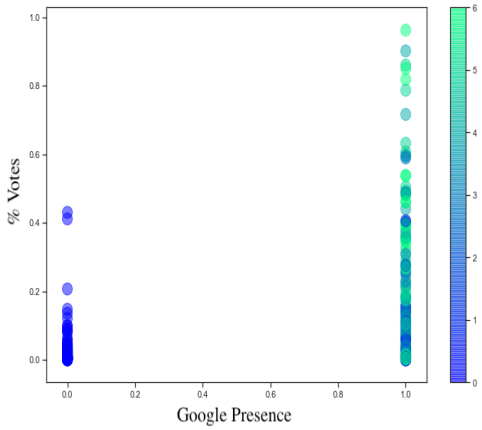

Fig. 5: Scatter plot of Google presence vs Percentage of Votes

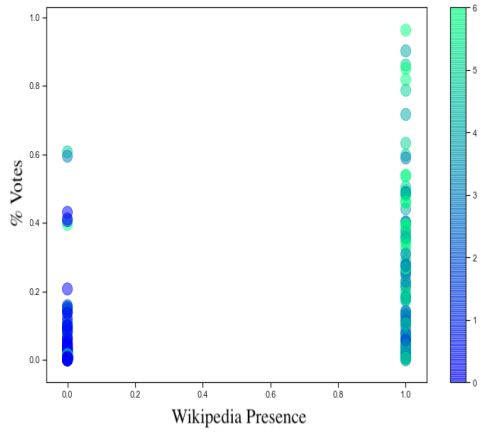

Fig. 6: Scatter plot of Wikipedia presence vs Percentage of Votes

4 Data modeling and Evaluation

Random Forest, also referred to as random decision forest, is best understood as a supervised learning algorithm. As the name suggests, it creates a forest and makes it random in one way or another. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result as shown in Fig.7. To put it simply, a random forest creates or develops decision trees and merges them to arrive at a prediction that is both more accurate and stable.

Table 1: Table captions should be placed above the tables.

Stats	Votes	Wiki	Google	No:Orgs	No:Events	No:Articles
count	244	244	244	244	244	244
mean	0.131	0.422	0.627	0.897	1.368	10.520
std	0.196	10	10	10	10	10
min	0.000001	10	10	10	10	10
25 percent	0.008	10	10	10	10	10
50 percent	0.033	10	10	10	10	10
75 percent	0.174	10	10	10	10	10
max	0.961	10	10	10	10	10

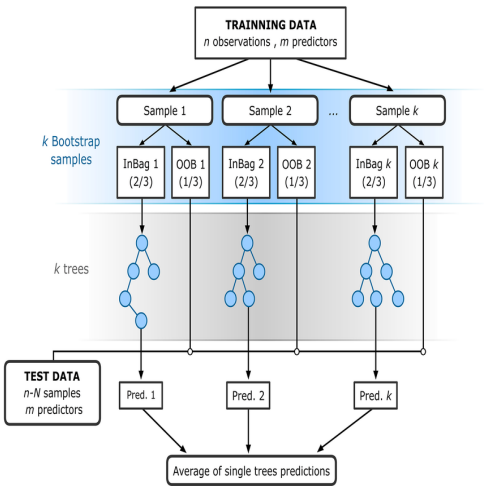

Fig. 7: Scatter plot of No: of Orgs vs Percentage of Votes - Articles

One notable positive aspect of random forest is the fact that it can be utilized in a range of uses including regression and classification problems, which typically make up most of machine learning systems used today[4].

Based on the above analysis and data exploration, it has been determined that a model can be formulated using the five features as shown in Table 1 with percentage of Votes as the output variable. Descriptive statistics for the features identified above, and for all states and candidates are shown in Table 1.

As a first iteration, Random Forest Regressor with 150 estimators, yielded a Pearson correlation of 0.5493 and MSE of 0.0355. The data has been divided to have an 50/50 split between training and testing data.

Further, new features such as no: of opponents and party affiliation of the candidate have been added to the Random Forest model with cross validation(cv=5). The second iteration of the model resulted in a Pearson correlation of 0.79 with MSE of 0.02.

Although second iteration was an improved model from iteration 1, the prediction of high performing candidates seemed to have higher deviation from the actual values.

This prompted for addition of new features that have been derived based on words with highest(top 10) Term-Frequencies from the Wikipedia pages of candidates with greater than 50 percent of votes.**tf-idf** scores of the above identified top frequency words, as shown in Table 2 are fed to the model as feature values. These features are added to the Random Forest model in addition to the ones from the previous models. This new model with 5-fold cross-validation on the subset of candidates with Wikipedia page(103 candidates) yielded a Pearson Correlation of 0.709 and MSE of 0.019.

Table 2: Additional Features - based on Term Frequencies(NLP)

congress	president
republican	bill
election	act
house	committee
senator	senate

Table 3 shows the various models used and the corresponding parameters optimized for each model.

Table 3: Random Forest Model Evaluation

Iteration	Data split	no: of estimators	MSE	Pearson Correlation
1	Train-Test: 50-50	150	0.0355	0.5493
2	Cross-validation: 5 fold	150	0.02	0.79
3	Cross-validation: 5 fold	150	0.019	0.71

5 Results and Analysis

Based on the above modeling and evaluation, for iteration 1, as depicted in Fig.8., it shows the predicted percentage of votes against the actual given by Random Forest Regressor. It can also be noted that the prediction percentage of votes are mostly being flattened at 30 percent. Although some of the predictions fall on or slightly away from the 45 degree line, a considerable amount of them are not in alignment with the Actual percentage of votes (away from the 45 degree line).

As highlighted in the previous section, for iteration 2, after including new features such as the number of opponents and Party affiliation, as shown in Fig.9, the prediction accuracy has improved in comparison to the previous model. However, candidates with very high percentage of Actual votes (greater than 80 percent) are not in alignment with the predictions (away from the 45 degree line).

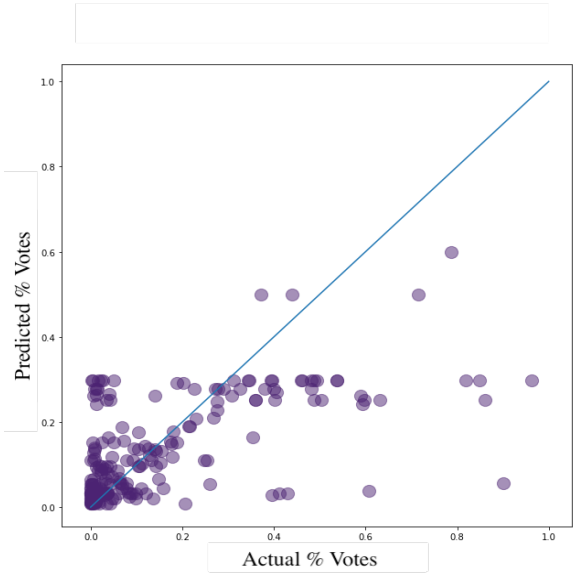

Fig. 8: Scatter plot of No: of Orgs vs Percentage of Votes - Articles

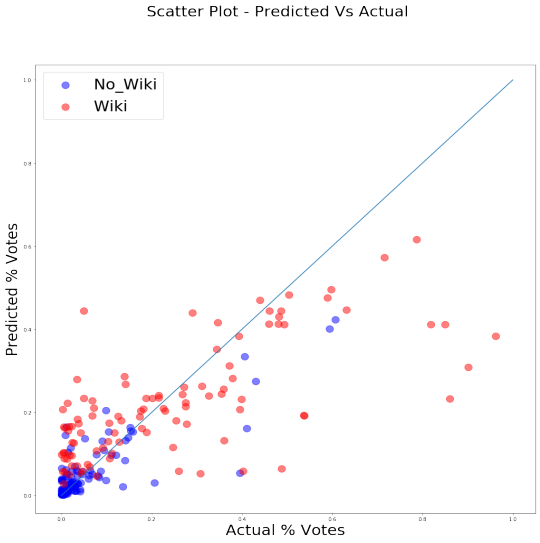

Fig. 9: Scatter plot of No: of Orgs vs Percentage of Votes - Articles

As shown in Fig.9, it is noticed that majority of these candidates with high percentage of actual votes have a Wikipedia page associated with them. The red markings in the figure represent candidates with Wikipedia page associated with them while the blue markings represent candidates without a Wikipedia page associated with them. Only a few that do not have a Wikipedia page associated with them identified with a

higher percentage of votes, a result that may be associated with other factors that may influence their popularity among the electorate.

For iteration 3, after including NLP features for the subset of candidates with Wikipedia page(103 total) and as shown in Table 2, the results of Predicted vs Actual votes are shown in Fig.10. Predictions for candidates with very high percentage of Actual votes(greater than 80 percent) improved in comparison to the previous iteration. In addition, majority of the predictions are very close to the actual percentage of votes.

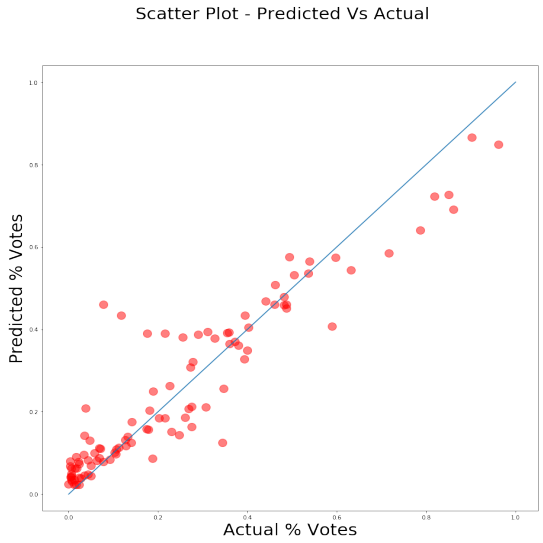

Fig. 10: Scatter plot of No: of Orgs vs Percentage of Votes - Articles

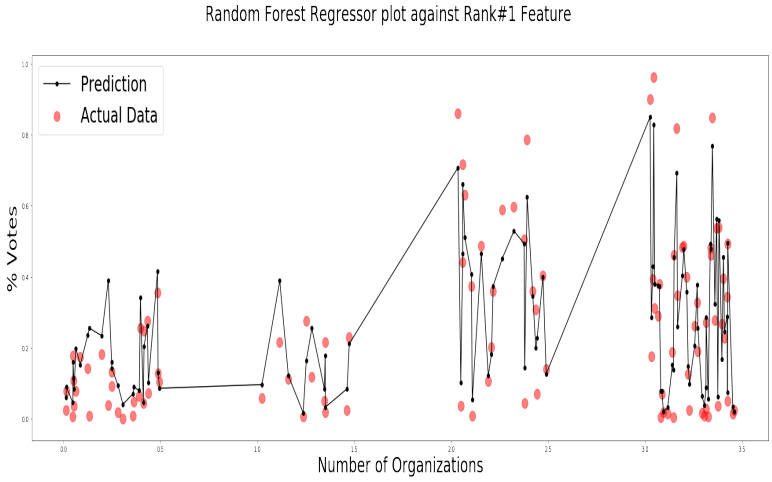

Fig. 11: Random Forest Regressor - Prediction vs Rank 1 feature

Following the deduction that this model holds well, the Random Forest Regressor prediction was plotted against the actual votes as indicated in Fig.11. The proximity

of prediction in black line to the actual votes in red dots further validates the model visually.

The prediction data between the values of 2 and 4 for number of organizations is further zoomed-in and it indicates the model's flexibility in handling the complex scenarios between those data points as shown in Fig.12.

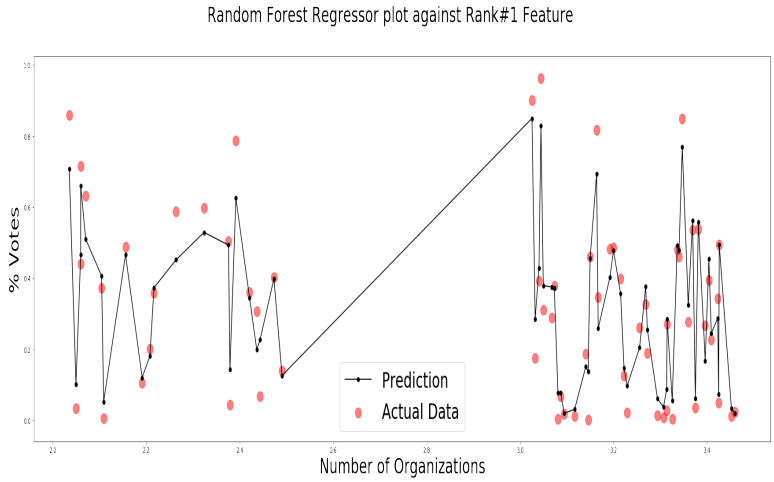

Fig. 12: Random Forest Regressor - Prediction vs Rank 1 feature

The final model is an ensemble model, that has both sets of data (candidates with/without Wikipedia pages) and the associated features as shown in Fig.13. It is noted that the candidates with a higher percentage of votes are being predicted more accurately than the previous models due to the addition of NLP features. Fig.9 shows the candidates with no Wikipedia data associated with them tend to have lower percentage of votes.

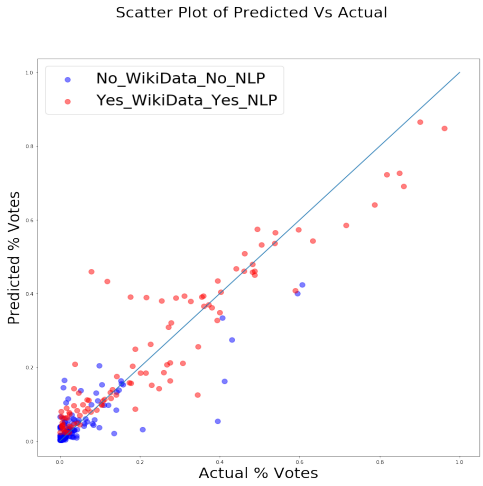

Fig. 13: Scatter plot of No: of Orgs vs Percentage of Votes - Articles

It is also noted that the 5 candidates(red tail on the left) with lower percentage of Actual votes are not being predicted well. However, these results were not consistent and changed over various iterations of 5-fold cross validation of the Random Forest model, while others being consistent. This is an indication of limited amount of data. But this variation is seen only in less than 5 percent of the data points.

Feature ranking has been calculated using Gini scores[9] as shown in Fig.14. This indicates that '**Number of organizations**' and other NLP key word features such as **congress**, **republican**, **house**, **bill** and **election** together explains 80 percent variation in the percentage of votes (response).

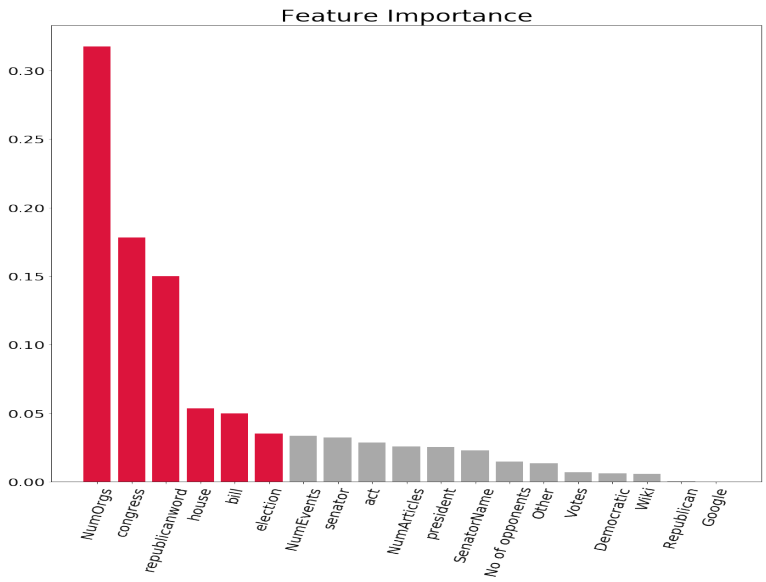

Fig. 14: Scatter plot of No: of Orgs vs Percentage of Votes - Articles

The final set of features are as indicated in Table 4(Features for base model without NLP) and Table 5(Features for base model with NLP).

Table 4: Base Features - without NLP

Google	No: of Articles
Wiki	No: of Opponents
No: of Events	Party
No: of Organizations	

[9] https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3

Table 5: Additional Features - with NLP

congress	president
republicanword	bill
election	act
house	committee
senator	senate

6 Ethics

Several ethical considerations have been made in this analysis. To start with, the three data sources used for our project, namely Google Knowledge Graph, Wikipedia, and News API are public and open data sources that are free to access to users and may be verified as needed. In addition none of them require authentication or approval for usage by any entity or authority and for this reason, no copyright or licensing issues may be attached.

Additionally, public data from social media platforms that might have legal implications has not been considered for our analysis. Lastly, data anonymity has been maintained throughout the analysis as well as its documentation for the purposes of privacy protection.

7 Conclusions and Future Work

In sum, our final model is an ensemble of two Random Forest Regressor models that identify with different feature sets as shown in Table 4 and Table 5. The feature set used in both cases depends on whether a candidate has a Wikipedia page associated with them (Table 4 and Table 5) or not (Table 4).

From the analysis, it is notable that political candidates that have a Wikipedia page associated with them have better chances of getting more votes in political competitions than their counterparts with no Wikipedia presence, which gives makes them more likely to win in political competitions. An 80 percent of variation in response is determined by the identified 6 features Number of organizations and other NLP key word features such as congress, republican, house, bill and election as shown in Fig 14.

The analysis pointed to a correlation between the percentage of votes and derived features. For this reason, the outcomes of elections are, in one way or another, dependent on a range of elements such as the affiliation of a candidate to featured organizations, their presence in public events as well as other factors like political background among others.

This analysis can be applied in a range of scenarios. To begin with, it can be extended to take user input and be converted to a recommendation system for indecisive voters. Secondly, it can also be extended to predict other election scenarios, which may be affected by the overall popularity of a candidate and the amount of information about them available to the voters.

Further, this analysis stands as a basis for future work that may be carried for the purposes of providing more insight on the political process. For instance, new NLP

features based on news articles can be engineered for the extended model that might have much deeper understanding using sentiment analysis of the content, thus deriving insights into the stands that different candidates take on various open issues.

Acknowledgements The authors would like to thank everyone who made the completion of this project possible. We are especially indebted to Ginger Holt(Facebook), who provided the group with extensive guidance on feature engineering.

References

1. Robert M. Groves and Lars E. Lyberg, An Overview of Non-response Issues in Telephone Surveys, in Robert Groves et al., eds., Telephone Survey Methodology (New York: John Wiley & Sons, 1988), pp. 203-205.
2. H.Paulheim. Knowledge Graph Refinement: A Surcery of Approaches and Evaluation Methods, Semantic Web Journal,(Preprint):1-20, 2016.
3. Wikipedia contributors. (2018, November 7). United States Senate elections, 2016. In Wikipedia, The Free Encyclopedia. Retrieved 05:50, November 7, 2018, from https://en.wikipedia.org/w/index.php?title=United_States_Senate_elections,_2016&oldid=867636691
4. Rodriguez-Galiano, Victor & Snchez Castillo, Manuel & Dash, Jadunandan & Atkinson, Peter & Ojeda-Zujar, Jose. (2016). Modelling interannual variation in the spring and autumn land surface phenology of the European forest. Biogeosciences. 13. 3305-3317. 10.5194/bg-13-3305-2016.
5. Bayesian analysis for political research Annual Review of Political Science, Vol. 7, No. 1., 483 by Simon Jackman