Political Profiling using feature engineering and NLP to help indecisive voters

Chiranjeevi Mallavarapu, Ramya Mandava, Sabitri KC, Ginger Holt

September 2018

Abstract

We are in a digital world where information is powerful. The technology has achieved great progress in the last decade in terms of how data is collected publicly and stored than ever before. The branch of public data has been key for influencing lot of successful businesses, public event outcomes and even government initiatives these days. With the data availability also comes the challenge of using it smartly and accurately.

We have seen lately that the choices of indecisive voters have been playing pivotal role in some of the political outcomes. So we have decided to focus on public data in the political area, which has been a famous topic recently. The information is constantly generated about politicians and is available digitally on the internet. Due to the breadth of technological enhancements there are lot of sources for it based on the way it is captured, for example blogs, public pages, news etc. So for a normal reader it poses challenge to read through all this information that is generated frequently and make some sort of sense based on that especially if it was an indecisive voter. So having a way to synthesize this information and presenting in a format that is easy and quick to understand will help these voters make informed decision and thus could lead to a better outcome.

In this project, we use data science techniques to collect information about politicians that ran for US Senate in 2016, clean and structure this data, derive features and model the data to come up with a political profile of the candidate. We will be mainly use feature engineering and Natural language processing(NLP) techniques to achieve this. The political profile of the candidate consists of features such as, total number of times the person is in news, frequency of positive or negative sentiment, total political experience in years, total number of twitter followers, education/work background etc. Based on these features we will come up with a “popularity index”. We will compare this popularity index to historical votes of these candidates to come up the accuracy of our model.

1 Introduction

1.1 Motivation

We are excited to take up the mostly talked about topic of the year, role of public data in political outcome. Ever since internet evolved into world wide web, the available information about people has been increasingly available publicly and is on the rise especially in political arena. We have been seeing lately how internet, thus public media forums, news, blogs play a key role in influencing political outcomes.

Having the data publicly available more freely has both advantages and disadvantages. Advantage is that we have access to this information easily in digital format which was historically not available with similar ease. One has to go to an encyclopedia or a local newspaper or even inquire in social circles. Disadvantage is that, it poses challenges when having lots of data, what to look at and how to derive meaningful insights from it and ultimately draw the right inference. This initiated the thought process for our project which is further continued below.

1.2 Problem Statement

In current day politics, blogs are written and events are held and news is broadcasted every day about politicians and this information is generated constantly on the internet. It is unwieldy for public to read through this information to make sense of a politician's profile or in other words what are these politicians like? This is especially important if they are indecisive voters and they either lack interest or time to sift through this information and might make wrong and uninformed decisions. Automating profiling on politicians based on this ever-updating public data would make the decision making easy for such indecisive voters. However, this needs feature engineering and NLP and it is not possible to do with naked eye.

1.3 Project Statement

To collect public data about politicians that ran for US Senate in 2016, synthesize and come up with a political profile and a "popularity index". Data is collected using Google Knowledge Graph API, Wikipedia API, News API, Everypolitian.org and Twitter. We will use feature engineering to define features that best influences the political popularity of a candidate. We will then use statistical modeling to come up with a model that best represents popularity as outcome using the features as model variables. Finally, we will use this model as a proof of concept(POC) for candidate scoring based on politician's names as input one provides for a given constituency.

1.4 Results

We have collected the data for CA senate political candidates that ran for US Senate in 2016, using the 3 data sources mentioned above. We came up with few generic features such as number of notable organizations a candidate is associated with etc.; that were noted to be influential on the number of votes for the candidate. We then visualized that data to validate our understanding of the most influential features that determined the "percentage of votes" of a candidate. We then used these as our features on the train data to come up with a model that gives "popularity index" of a candidate using Random forest regression method. We used 2-fold cross-validation to determine the best number of sample groups. In the next revision, we plan to run this through few more Machine Learning methods to get the best method.

Based on our assessment of accuracy of this model output i.e. percentage of predicted votes against percentage of actual votes, Random Forest Regression method with 2 folds gave us the accuracy of 93 percent, so we decided to use this on test data set of candidates running for Nov 2018 senate elections to predict the outcomes.

We also plan to apply natural language processing on some of the free text from News API, Google KG and Wikipedia data to come up with sample features that are of interest.

1.5 Conclusion

Based on our data collection on politician's biographic data, public pages, news and blogs, it seems that the percentage of votes for a candidate are influenced by a number of notable organizations a person is associated with, thus indicating an active involvement in the community and the number of events that a politician is part of, indicating his/her social influence and also his/her participation in both private and public news media giving them a good reach to the public. Although, our results indicated a 93 percent accuracy, we believe it is currently over-fitting due the lack of data across all states.

Some interesting points that our model gave us to ponder about:

- Candidate: Ron Unz, where the predicted votes exceeded the actual by 17 percent based on our feature selection, which meant that he has an active presence in the community. Upon further analysis we noted that he decided to run for 2016 senate elections in the last minute which might not have had a good reach to the CA population of voters

- Candidates with highest votes (Kamala Harris and Loretta Sanchez) have a much stronger positive correlation to their associated number of organizations and events and this can be explained due their minority background and CA being a majorly minority populated area.

- All candidates that have 20 percent or more votes either have non-zero events or non-zero articles, suggesting that those features are influential in those cases. The only case where this was not true is where we see percentage of votes is 7.77 and the candidate is Duf Sundheim. It is also observed that this candidate has a political background.

We further plan to refine our features to make the above model more accurate. We also plan to add data for the rest of the states and test our model for accuracy and evaluate any other features we might have missed during this initial modeling.

2 Related work in this area

We found that a similar work was done at vote-usa.org where one can choose their state and dig into the candidate running for senate and get high level information on their bio page.

However we noticed that this information is static and more text based one still has to read through to understand. Also it has a lot of subsequent links one has to browse through to get a sense of the political leader[9].

3 Data sources aligned with the problem

Based on our research, we identified five relevant sources of data about political candidates, and they are Google, Wikipedia/Wiki-data, Everypolitician.org, Twitter and Newsapi. Some other sources that we considered in the initial evaluation are Facebook and Linked-In. Although they can provide good amount of information about a politician's social media presence, due to the technical API limitations of these social media platforms in the recent wake of privacy issues, we chose to go to mostly used public sources. However, considering the presence of the candidates on social media platforms has influence in their popularity index, we are considering to gather twitter information on the candidates.

We will be using Google's knowledge graph API, which is an enhanced version of Google normal search function. A knowledge graph uses standard entity relationships between the information and structures it in such a way that one can make sense of it. For example, if one is searching for organizations associated with Donald Trump, the knowledge graph is smart enough to show not only the original presidential transition into Whitehouse but also all other organizations that Donald trump has been historically part of e.x. Trump Productions (TV production company), Trump Organization (Real Estate Company). So, we will be using this functionality of Google's knowledge graph to gather relative data points for politicians.

We will also be using Wikipedia's API to gather information about politician's bio and other relevant items particularly political history. Wikipedia being a publicly maintained data source we consider this being unbiased source of data about politicians.

Everypolitician.org is another amazing source of every politician's data from 233 countries around the world. It aggregates the data from different online sources, like official parliament sites and unofficial sites, merge and publish them in JSON and CSV format. The data contains names, dates of birth, twitter handles, political group memberships, email addresses, honorific titles, image URLs, and all sorts of useful things of every politician which is constantly monitored, updated and contributed by people around the world.

We will also consider News API from https://newsapi.org/ 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.

We plan to get the data also from twitter API for the identified senators to understand their interactions with the general public using sentiment analysis.

4 Data collection methodology

4.1 Google's Knowledge graph

Google's Knowledge graph 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 https://schema.org/ nomenclature. We have identified following key schema types in relation to politicians that we will be using : Person, Event, Organization, Book

4.2 Wikipedia

Wikipedia rest API along with pywikibot will be 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 Wikipedia's information about a politician is divided into biography and personal section, political career section, pywikibot gives the necessary tools to make knowledge graphs. However, we will be using the API to get this basic info about the politician to use along with google to come up with a profile.

4.3 2016 candidate Wikipedia pages

We will also collect the data about the politician's success rate from Wikipedia 2016 senator page using JSON download to get the total number of votes they received as well as their election outcomes to evaluate the accuracy of our features.

4.4 Everypolitician.org

We are also pulling basic bio and social media information about politicians from Everypolitician.org in a CSV format. This data source is a part of poplus.org initiative which is an international organization for using technology to enable civic use cases. We will use the flat file downloaded from here to identify the biographic information about the politicians and also their twitter page info.

4.5 Newsapi.org

We gather information about the latest news from Newsapi.org using their API about these politicians that gives us all relevant information that is in the news lately. This helps us derive some useful insights using NLP about a politicians activity and influence in public forums.

4.6 Twitter

We are currently evaluating possibility of adding Twitter feed for these politicians from their public pages. [TBD]

5 Data consolidation and accuracy

We would like to start with the current list of senators ran for 2016 as our data to model and run through Google KG [Person, Organization entities], Wikipedia APIs [bio, political history] and Everypolitian.org[bio]. Once we collect the basic bio data about these senators using the above mentioned APIs', we will then consolidate it in a structured way. We then collect the data from NewsAPI.org , Google KG[Events] and Twitter which gives us more text based information about these politician to data mine. We will then use Natural Language Processing techniques to come up with few features that best represents these varied data sources, for example number of events particular politician attended or number of news events about politicians that have a positive sentiment etc, explored further in detail below.

Once we derive the above basic bio data along with derived features we explore the data to check 1) data completeness 2) accuracy of data 3) quality of the data.

We started our exploration with CA as our test(focus) state and our results indicated that for the 37 candidates that ran for the senate, we were able to get 7 matches from Wikipedia and x form google and y from News API. We then created a consolidated data set that contains bio-data from Wikipedia and person/organization/event entity information from Google knowledge graph, and news articles form News API. The search results for the candidates were not 100 percent, as we could only get 20 to 30 percent of standardized data for these candidates and this was something that we expected at the beginning of this project due to the early stages of internet data standardization in general. So, we proceeded with this data set keeping in mind that this effort generates only a subset of data-set but could fully represent all aspects of our anticipated modeling.

Once the data is consolidated in this format, we further approached this problem by defining certain features that are derived from the above varying data sources. Based on our conversations with subject matter experts in the political domain, we included features such as number of events a person might have attended, number of organizations a person is associated with and number of news article about a given person. With these features defined, we went ahead and visually explored the data in fig 1.

It is apparent that the percentage votes vary based on the number of events and organizations a candidate is associated with. However, there is no straight forward linear relationship. We see two different patterns here: One that would not change the votes based on increase in his/her number of news articles, another where we see a linear relationship between the number of articles and votes.

As observed in fig 1, candidates with highest votes (Kamala Harris and Loretta Sanchez) have a strong positive correlation to their associated number of organizations and events. However, this data is skewing rest of the information due to a great majority in the percentage of votes section. This pattern in data is understandable considering the demographic information of the top two candidates (Kamala Harris and Loretta Sanchez) being influential in determining their success, in relation to the state (CA) they competed in. So, we would like to remove these data points to further explore the other candidates and their relationship to the engineered features.

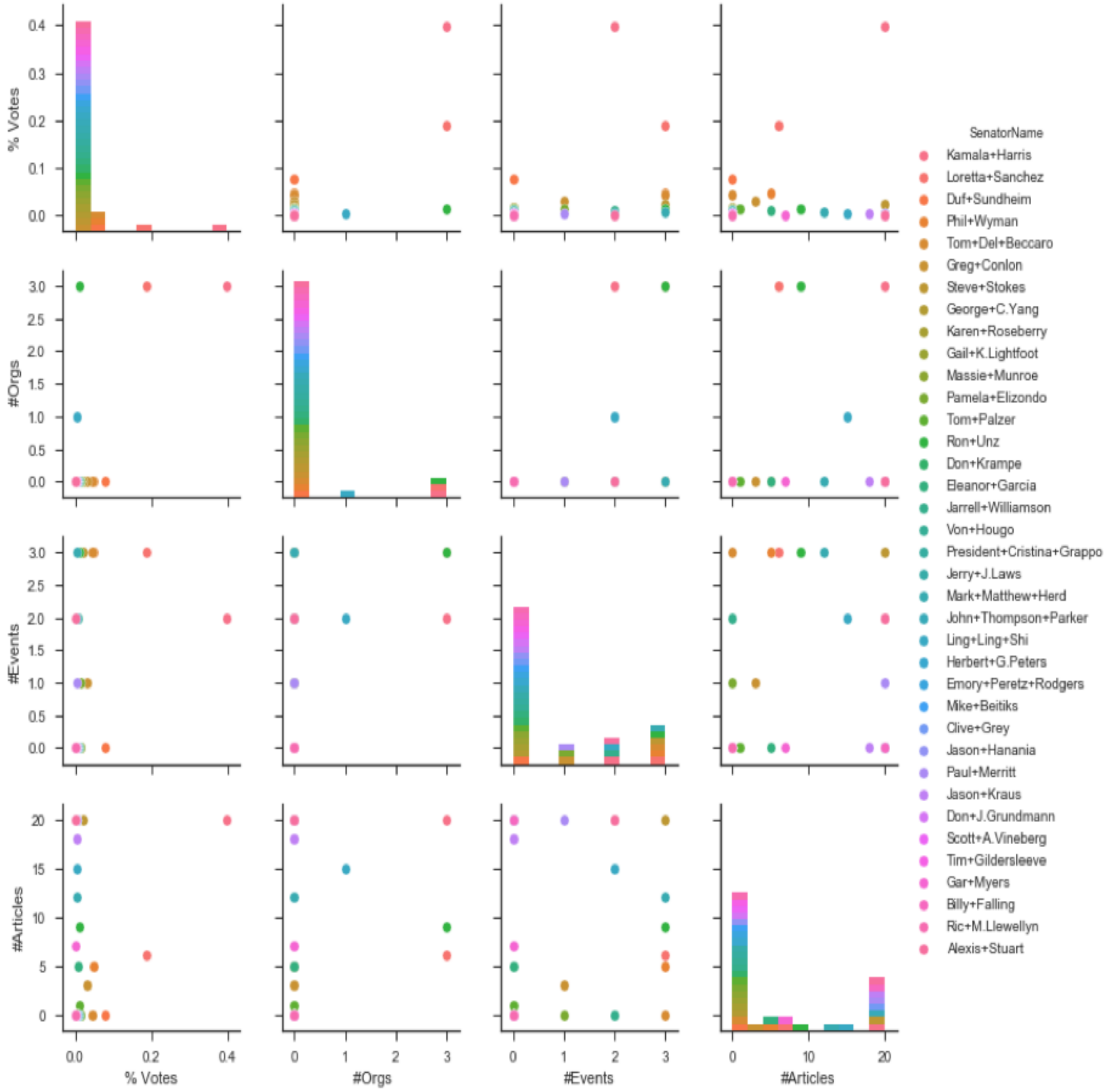


Figure 1: *Pair plot of engineered features*

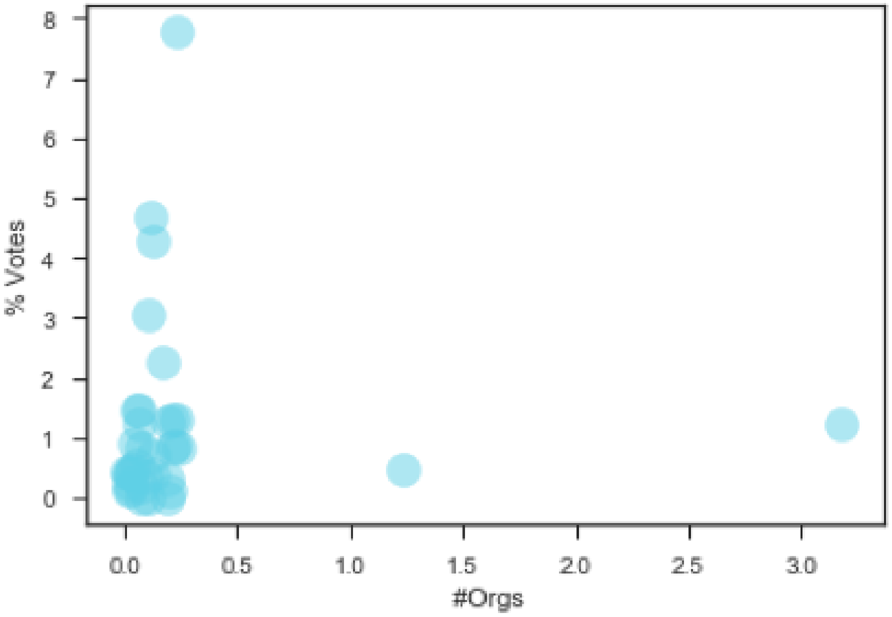


Figure 2: *Pair plot of engineered features*

5.1 Percentage of votes vs Orgs

As observed in fig 2, after excluding the top two candidates (based on votes), it is that only two other candidates have association with organizations(non-zero), and it looks like this association has an influence in gaining vote base. It is also interesting to see that the people that do not have any generic association with an organization still gained good votes that we need to explore further.

In order to understand the cluster of candidates that have non-zero votes with zero association with organizations, we included a color scheme based on number of Event and Articles that they are associated with as depicted in figures 3 and 4. It is evident from these plots that, all candidates that have 20 percent or more votes either have non-zero events or non-zero articles, suggesting that those features are influential in those cases. The only case where this was not true is where we see percentage of votes is 7.77 and the candidate is Duf Sundheim. It is also observed that this candidate has a political background.

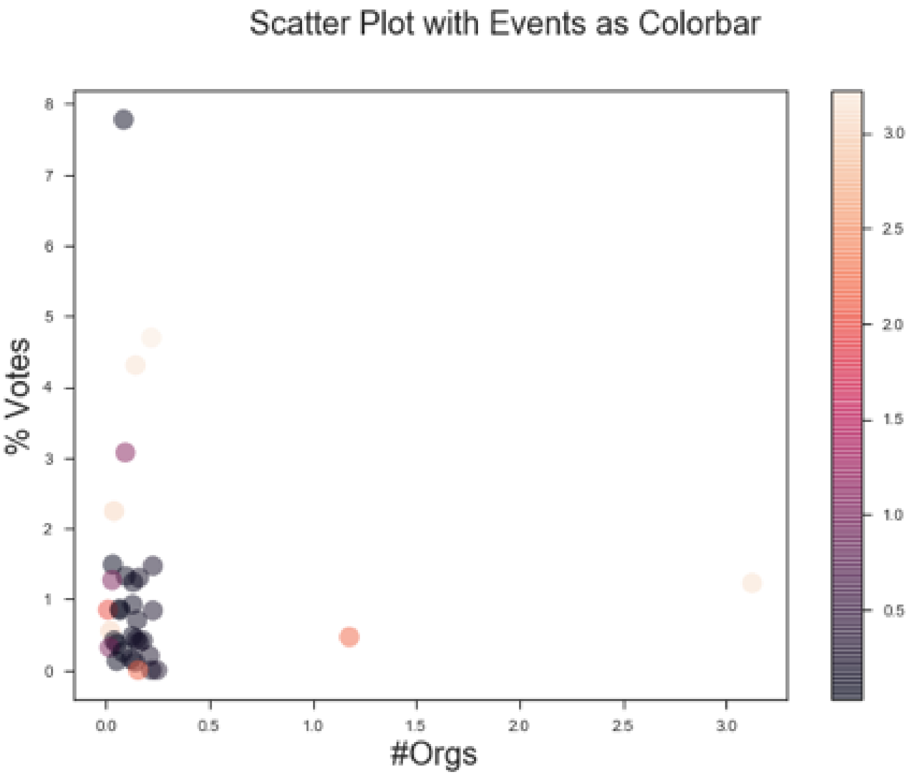


Figure 3: *Scatter plot of Orgs vs. Percentage of Votes - Events*

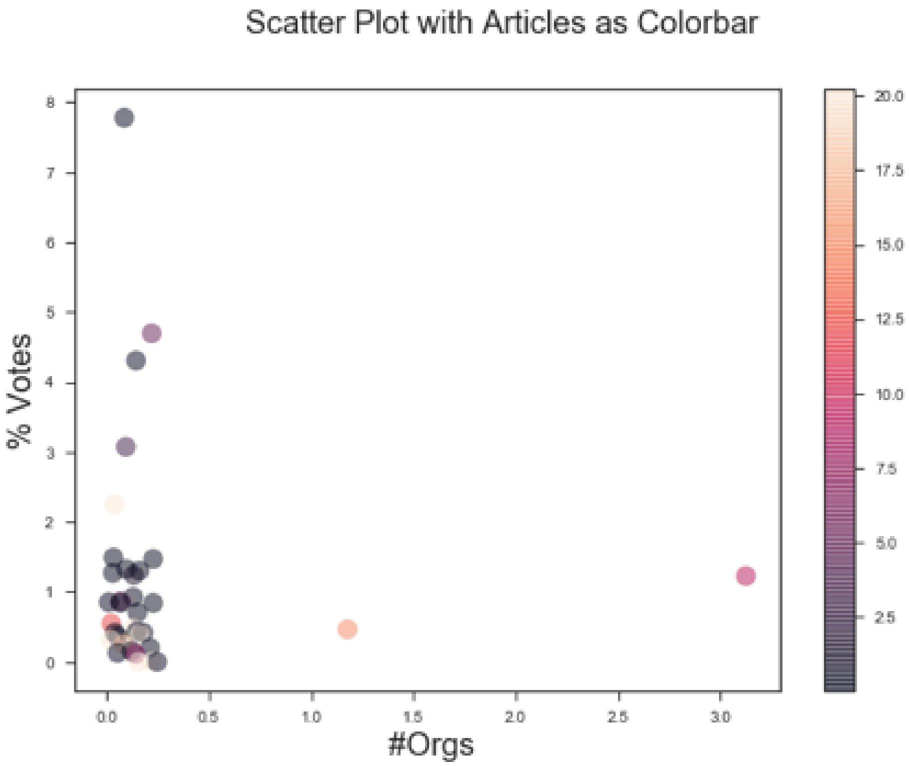


Figure 4: *Scatter plot of Orgs vs Percentage of Votes - Articles*

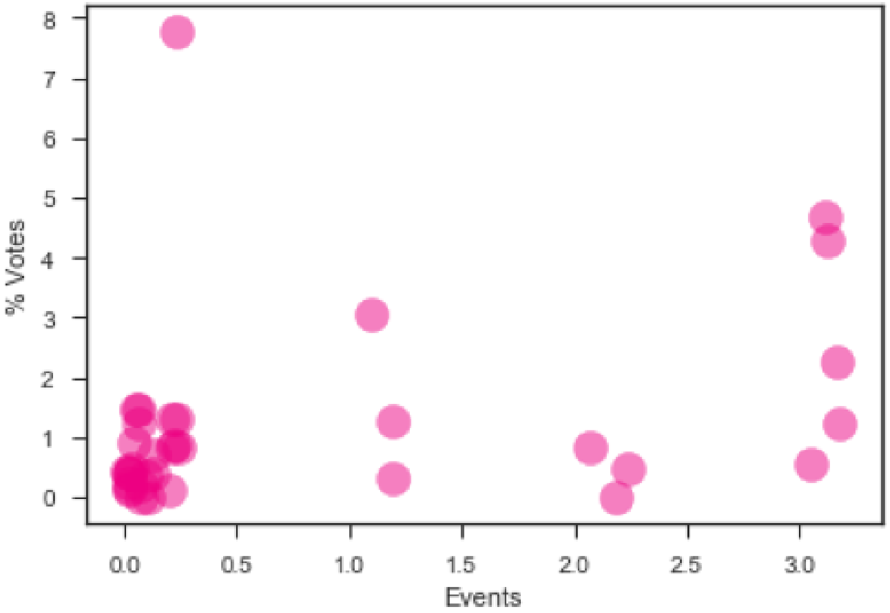


Figure 5: *Scatter plot of Events vs Percentage of Votes*

5.2 Percentage of votes vs Events

A clear positive correlation can be observed between the number of events and percentage of votes from fig 5. The one candidate with 8 percent of votes and zero events seems to be an outlier (Duf Sundheim).

5.3 Percentage of votes vs Articles

As can be seen in fig 6, the one outlier with 7 percent votes is Duf Sundheim, whom we discussed in section 5.1. Otherwise, there is a general positive correlation between number of articles and percentage votes. However, for all the candidates where we noticed the number of articles are more than 7.5 but with small percentage of votes, it is possible that the articles are not an exact match to the person we are searching for. This prompted us to further validate the accuracy of these articles from news API using NLP techniques which we will explore for the next revision of this document.

We then take steps to clean or balance the data set with necessary adjustments using imputation techniques. Once we have a clean data set, we will be using data visualization techniques like scatter plots /clusters to further explore

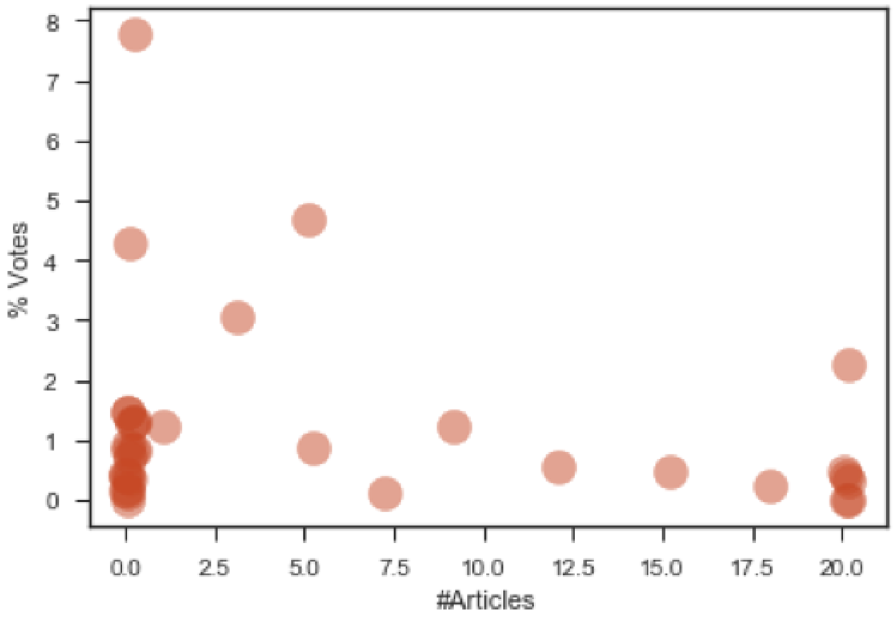


Figure 6: *Scatter plot of Articles vs Percentage of Votes*

the data to get a sense of relation of these derived features to number of votes these politicians achieved in 2016 senate elections.

We will then assess some of the possible features that can further define this data using feature engineering techniques. Part of this also involves features like whether a particular event has positive impact or negative impact on society, using NLP techniques. We will be counting the organizations that a particular politician is involved in, to get a sense of his/her background in the business world. We will also measure the frequency of these events to calculate the activeness of the candidate.

We plan to constantly evaluate and refine these features during the development phase as we further explore their impact on the outcome and hence, a moving target.

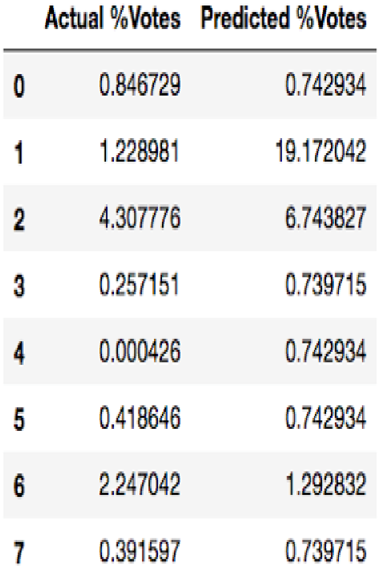
6 Data modeling and evaluation

The data extraction from generic data sources such as google, wikidata and newsapi would leave us with plenty of unstructured data to begin with. Instead of trying to make certain predictions immediately, we would like to use k-means and hierarchical clustering methods for initial exploratory data mining and format our data to make it structured. Further, techniques such as tf-idf(term frequency-inverse document frequency) would be used to create quantitative features necessary for our model predictions. As shown in the below formula 1, one of the studies describes TF-IDF as,

> TF-IDF takes the product of the frequency of a term and the inverse log of the frequency of the number of documents containing the term within a collection of documents, where nt represents the number of times a term t occurs in a document with N total words in the document, dt represents the number of documents in a collection containing the term and D represents the number of documents in a collection. In this way, TF-IDF down-weights term frequency within a target document by how many times it appears in other documents in the collection, thus scoring terms favorably for being both frequent and unique within a given document(Stoffa, Lisbona, Farrar Martos, 2018, p. 4).

$$TF-IDF = nt/N.(log(dt/D))^{-1}$$

$$Formula(1)$$



	Actual %Votes	Predicted %Votes
0	0.846729	0.742934
1	1.228981	19.172042
2	4.307776	6.743827
3	0.257151	0.739715
4	0.000426	0.742934
5	0.418646	0.742934
6	2.247042	1.292832
7	0.391597	0.739715

Figure 7: *Random Forest Regression: Actual vs. Predicted*

Based on above data that is formatted in a structured way and has features derived, we will be using standard supervised classification methods like Decision trees, Random forests, Support vector machines, Neural networks and Bayesian classification to come up with the best classification model.

Based on the feature engineering that we did, we determined the most influential features are SenatorName, association with number of notable organizations, number of prominent events a candidate is associated with, number of news articles a candidate is mentioned about, if the candidate has a Wikipedia page, and if the candidate is optimized for google search engine (assuming this indicates his popularity).

Using this feature set, we used the random forest regressor to come up with a model to predict the candidate votes. The model yielded an accuracy of 93 percent with MSE = .00412 and Pearson coefficient = 0.3(out of bag R square = 0.53). The data has been divided to have an 80/20 split between training and testing data. We plan to further improve the model through cross validations and hyper parameter optimization, as well as we further refine our features.

Table 6 shows the predicted percentage of votes against the actual given by our Random forest regressor. It can be noted that both values are close except for row 1(Candidate: Ron Unz), where the predicted votes exceeded the actual by 17 percent due to the high number of feature values we engineered. Upon

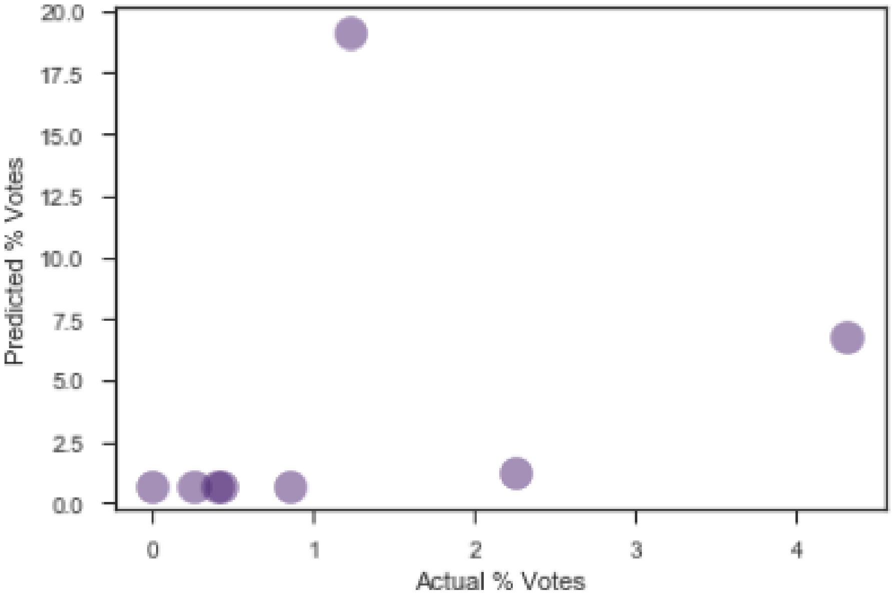


Figure 8: *Random Forest Regression: Actual vs. Predicted scatter plot*

further analysis we noted that he decided to run for 2016 senate elections in the last minute which might not have had a good reach to the CA population of voters [15].

The output of our model on test data set of 2016 political candidates will be "popularity index" (percentage of votes received vs total voters in the geographical area). We will be using the candidate's prior voting/success history to determine the "popularity index" for test data set and then the impact of the features on this outcome to measure the accuracy of our model. We will be using multi-fold cross-validation techniques to eliminate over-fitting of the model.

Once a model is determined, we can pass few local candidates for a given constituency and the model will provide scores based on the features identified so far along with an accuracy measure to show how good or bad the score is so the end user has a sense of how to apply these scores in their decision to vote for a particular candidate.

7 Data Application POC

Once we finalize our model, we would like to use this model primarily for indecisive voters to make an informed decision. One can use our tool to input few politician names that are running for local senate or even presidency and the model gives out with an overall score for each candidate with a detailed profile about some of the features that best influence the popularity of the politicians. It also gives all the relevant articles that are related to a particular feature if one chooses to further explore in detail. Our model also provides an accuracy score so it helps people to make decision based on their best judgment.

This helps lot of indecisive voters or even voters with a prior choice to look at the information in a different lens and possibly change their decision to a data-driven decision rather than 'gut' feeling.

8 Further exploration, Industry Applications - looking ahead

Lately it has been noticed that media like TV, Radio, Newspapers have been using information or better put "misinformation" to influence public events like presidential elections or even BOX office results of a movie. So the information available on the web about a public figure like a politician or a celebrity whether it is Wikipedia/social media/news paper article/blog is growing its importance every day as long as it is not controlled by biased media entities but rather by entities of public interest.

Even political voting outcomes are measured by the pulse of what people like to talk about on public forums these days. So our project would like to solve this problem of "misinformation" influence by providing a tool that can gather public information about a politician and synthesize this information to provide a more unbiased view as a profile of the public figure. We anticipate a tool like this will help people to get better awareness of the politicians they are voting for.

In addition to the "popularity index" we would like to also include donation history of the candidate and which groups with special interests support such donations to really dig further into associations with other organizations.

We also strongly believe that "social influence" also plays a role in choosing political choice hence we would like to extend this into an app where people not only have visibility of their own political choices but an anonymized political preference of their friends and families and other inner circles.

With this 360 degree view, the long-term aim, and future exploration of our project can be extended to state and local level or even worldwide election candidate profiling and help people to increase awareness by stripping down the complexity of election and choosing the right candidates. It will also help people get freedom of information and develop transparency through digital platforms. With a concise and transparent information, it will also help to eliminate the biased medias and yellow journalism against politics and politicians.

References

[1] http://journals.sagepub.com/doi/pdf/10.1177/2053951716645828

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

[3] https://en.wikipedia.org/wiki/United$_{S}tates_{S}enate_{e}lections,_{2}016$

[4]$https://www.quora.com/What - are - the - free - news - feed - APIs$

[5]$https://newsapi.org/docs/get - started$

[6]$http://www.euclid - project.eu/iBook$

[7]$http://knowledgegraph.info/$

[8]$https://everypolitician.org/united{-}states{-}of{-}america/senate/download.html$

[9]$https://vote - ca.org/Intro.aspx?State = CAId = CAHarrisKamalaD$

[10]$https://www.mysociety.org/files/2014/11/FOI{-}Impact_{P}art{-}2{-}Practitioner{-}Study - 06.pdf$

[11]$https://newsapi.org/docs$

[12]$http://docs.everypolitician.org/technical.html$

[13]$http://docs.everypolitician.org/repo_{s}tructure.html$

[14]$Stoffa, J., Lisbona, R., Farrar, C., Martos, M.(2018). Retrievedfromhttps://scholar.smu.edu/cgi/viewcontent.cgi?article = 1005context = datasciencereview;$

[15]$http://www.unz.com/runz/a - last - minute - decision - to - enter - the - u - s - senate - race - in - california/$