Bootstrapping for polling data from the EU referendum vote in the United Kingdom*

Alexandros Georgios Stergiou[1]

Abstract— **Voting polls have been the most prominent and widely used way of forecasting results. The most straightforward way of determining the voter's intentions would be to ask who or what he or she is going to vote. However, it has been found that even how this questioned is phrased can effect the results. Ambivalent voters as well as vacillating voters who have not fully though their decision can have a significant impact in both the polling results as well as towards a possible divergence from the results during the voting day. For voters in these cases, addition information such as their features can signify the most probable tend for each individual in regards to similar examples that have very strong opinions. By considering these features, what is proposed is a simulation of each person's experience, when casting the ballot. This paper presents the use of *Bootstrapping* (or also known as bagging), to normalise the features of the polling population in order to represent the entire population that is eligible to vote. This technique will be applied to data from the 2016 Referendum that held the question of the remain or exit of the United Kingdom from the European Union.**

I. INTRODUCTION

The use of opinion polling to simulate election results has been a prediction method that has been used extensively. The first recorded use of polling was in 1844 in the Contest of the American Presidency between Andrew Jackson and John Quincy Adams. However it was until the early 20th century where opinion polling broke the barrier of local phenomena and was used nationally by *The Literary Digest* which correctly predicted Woodrow Wilson's victory in the Presidential election in that year[1]. Since, polling has been conducted by both government and private companies for estimating the voting outcomes. Howbeit, recent results have shown that in many cases, it is possible that outliers may occur in the sample that shape the polling results in a way that they do not represent accurate expectations for the event. This may be due to a variety of reasons raging from polling errors (such as using biased towards the population sample chosen[2]) as well as psychological reasons (people preferring to reserve their true answer). Past efforts mostly were targeted towards choosing the right sample that can best represent the entire voting population and draw immediate conclusions from the data received. Apart from the initial difficulty of this task, considering all the socioeconomic factors and features needed to to be compared in order to create a suitable sub-population by finding the appropriate candidates is infeasible in practice. For this reason, a different statistical method is required after the opinion poll has taken place.

Re-sampling has been a fruitful method on estimating the precision of the sample by drawing random examples from the data[3]. This allows the minimisation of errors based on outliers in the distribution, in population parameters such as the mean or the median. With the much increasing use of computational power, bootstrapping large sample has been made widespread. The idea that was mentioned uses the categorical demographic data that has been made available in the survey and compare them with the ones drawn by the entire population. Then by using bootstrap, new examples are generated by randomly re-sampling the data and finally, when reaching a satisfactory point (where the distribution resembles the one of the entire population), draw an estimate of the result. Although for most tasks it is important to know the true confidence interval, bootstrap's simplicity has been found to be more accurate[4] than other methods.

*This work was produced for the fulfillments of Data Science and Decision Making module *CE888*

[1]School of Computer Science and Electronic Engineering, University of Essex, United Kingdom

II. BACKGROUND

A. Similar Projects and Systems

One of the most important and widely accepted methods for predicting voting outcomes was created by Steven Rosenstone in 1979 in *"Forecasting Presidential elections"* [5]. The most important feature that was introduced by J. Rosenstone was the use of characteristics from each questioner in order to make a prediction. Although the final equation achieved a standard error of 4.5% and mean absolute error of 2.9% for the state vote, some parts may require further investigation. For example, many answers were determined by panels of scientists that assigned an appropriate value. These values may have been strongly related to the belief of each of the scientist in terms of how representative it is towards the answer given. Therefore, some of the values recorded may have been underestimated (an example of misleading data gathered was the coalition of different minority groups and labor workers that led to Truman's election in 1948). Also, the equation used data that were published after the event, for example the economic growth of the state for the period.
Although the J. Rosenstone equation was not optimal, it was the starting point for political forecasting as the subject gain more friction with the introduction of the prediction market a decade later at the Iowa Electronic Market[6]. The first subject that the market allowed to bet on was the 1988 US Presidency where the shares were only binary (one for each of the two candidates). The market was run by the University of Iowa business school and it had three "submarkets":

- The General Election market
- The Republican Nomination market
- The Democrat Nomination market

The purpose of this market was educational, as the researchers updated the rates with new data that were obtained. As the computation power increased, the methods used progressed. The two most dominating methods that were presented were:

- **Combining poll data and averaging them.** It has been shown that combining forecasts can reduce errors. The most significant advantage of this method is that the bias associated with polls are likely do be different in each experiment and therefore they will be minimal (or close to none based on the number of polls), in the averaged population. Furthermore, the combination of forecasts will provide additional information that allows to form a fuller picture of the influences that affect possible errors[7].
- **Poll damping.** This methodology does not consider incorrect indicators of public opinions[8]. It was used to estimate results of 12 presidential elections from 1948 to 1992 and achieved a mean absolute error of 1.33 %. For example, the first polls that are carried out are considered of poor measure in comparison to polls closer to the voting day.

B. Current Efforts and State of the Art

Most systems that have been developed for opinion polling today make use of the constant data flow in social networks such as Twitter. Estimates on people opinion's can be drawn from posts and captions that show not only the individuals political views, but also the emotion felt towards a a party or a candidate. Such systems were also used in the US 2016 elections where live tweets were fed to a system that did not only consider sentences where a name of a candidate or a party was mentioned, but in addition, translated the emotion felt by the voter's choice of emoticon(s) to associate with his/hers tweet[9]. A dictionary was used in which, five emotions could be represented with different emoticons (happiness, sadness, laughter, anger and fear). Then, three specified classifications where used to find the opinion of each person (Naive Bayes, SVM and Nearest Neighbour), with precisions raging between 45-50% and recall of approximately 50% for each candidate. The use of live generated information can also be utilised to find the trends and how statements and actions can affect the voter's choice. However, it is important to note that a positive comment towards a politician does not mean a guaranteed vote. In addition, a significantly demanding task is the recognition and the correct classification of tweets and posts as the context may be unknown or they might be made for the purpose of humor rather than a solid statement.

III. DATASET

The datasets used in this project are separated to two groups. The first group includes polling data and the second one has demographic information based on the 2011 census or private company survey data (that will be assumed to represent the tends of the general public).

The sample to be used was obtained by the YouGov website and contained two opinion polling data; one from online questions and one from over the phone. For the needs of the task and at the current state, all the results gathered were based on the online survey as different questions were asked at the second. Additionally, the online questionnaire contained a data size of 2000 respondents while the phone examples were only limited to half of that size. In total, each example in the data contain thirty features with some exceptions based on previous answers or questions that were not asked (for example if the person did not vote in the 2015 election there is no point for a question about the voting party that he/she voted for). Considering the large amount of features, it is more reasonable to only bootstrap new examples based on the most important features. These characteristics are:

- The age of each person in the dataset.
- The gender of the voter
- What individual education qualifications they have obtained, taking into account university degrees, diplomas, certificates, apprenticeships, etc.
- The social grade group that each person is assigned.
- The vote casted in the 2015 general election.
- What news station they prefer to watch.
- The newspaper that they buy or read more frequently.

For each one of these traits a designated numeric value has been assigned to represent a response. In order to make the data meaningful, an additional codebook has been provided with the definitions for each of the values. It should be mentioned that there is not a specific reason for choosing these features from the larger pool of thirty attributes apart from the fact that, they used for personal evaluation by many other systems and therefore they have the potential to be the most relevant, for the purposes of this study.

The other group of datasets that will be used is composed from different sources and is based on the population characteristics that will be studied and used for bootstrapping. Therefore there are six datasets:

- **2015 General Election results:** This dataset includes the previous election results and each party's percentages. The source that the data were composed from was the BBC website[10], however not all parties are included, but only the ones that were present in the online survey as well. The rest were classified under the "Other" label.
- **Education qualification demographics:** These are grouped to six distinct qualification levels based on the FHEQ framework. The data comes from the 2011 census of the National Statistics Office[11]. Among the percentages for each education level, the number of people is also displayed as well as the rank of each ministry's rank per level group.
- **Gender and age ratios:** The data in this file include the number of people and the gender ratio of each age[12]. In the project, each age is not considered separately, but as part of a group, with each containing four ages.
- **Social grades indexes:** These are the indicators provided by the NRS[13] that show the income and the financial state for a group of people.
- **Newspaper preference:** Considering survey estimates for the most popular newspapers that are being read in the UK[14].
- **Station ratings based on fondness:** These are the ratings of the most favourable TV news stations in the United kingdom based on 2008 BARB survey[15].

TABLE I: Datasets and their features table

Dataset	Number of features	Used features
YouGov	30	7
2015 Vote	2	1
FHEQ level	22	6
Gender/age	2	2
Soc. grade	5	4
Newspapers	3	1
TV News	1	1

IV. METHODOLOGY

As the title suggest, the main tool for re-sampling that will be used is Bootstrapping. The system created will use the online questionnaire as the initial dataset in which the experiments are based on. Then the rest of the datasets are loaded in the form of Panda's DataFrames and reconstructed, if they do not include percentages. Then, once all the data are comparable a function is called to compare the actual demographic data with the ones obtained by the survey. If the data are not equal, bootstrapping is used to generate new examples with the use of a hill climbing technique and with the combination of Nearest Neighbour approach, to assign the most probable voting value, instead of a randomised estimate based solely on bootstrapping. This means that in each stage a condition is used to find if the data is representative of the general population. So the condition can take three different values:

- **The features are found to be over-represented.** In this case nothing can be done to normalise the data as it is not desired to delete examples, since information will be lost in that case. Instead, the method will continue to the next characteristic, because if some features are found to be over-represented, consequently others will be found to be under-represented.
- **Attributes are found to be equal between the two sets.** This is the ideal situation that we desire. Once reached at this case, a flag is raised which will be validated in the next iteration to find if a satisfactory number of conditions are true for the new generated data.
- **The characteristics are under-represented.** The final case is that the survey data do not hold adequate data for a feature and therefore, it is required to bootstrap new examples. Some attributes however need to be treated with care, for example to reassure that a person that did not voted does not include a voting party or that an age falls within the age group and is not simply the top or bottom value of the group.

The system considered each feature separately and will move to the next feature after it has bootstrapped a sufficient amount of examples or

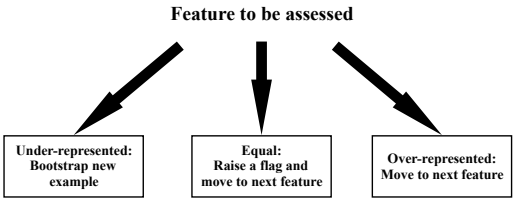

Fig. 1: Three different instances of the features

the number of times that the feature is present is equal to (or more than) the one desired. This process continues for every feature until all the characteristics are inspected and thus, it has finished the first main iteration circle. The number of circles is strongly defined by the distribution of the data, as a less similar to population distribution is expected to require more iteration circles that another with a higher similarity. The final step is to move through the next field. For example if the system bootstrapped all the needed examples for the age groups, it will next move to the social grade data and apply the same procedure. So hill climbing is used to bootstrap for both the individual attributes as well as the fields. A representation of the process can be seen in Figure 2.

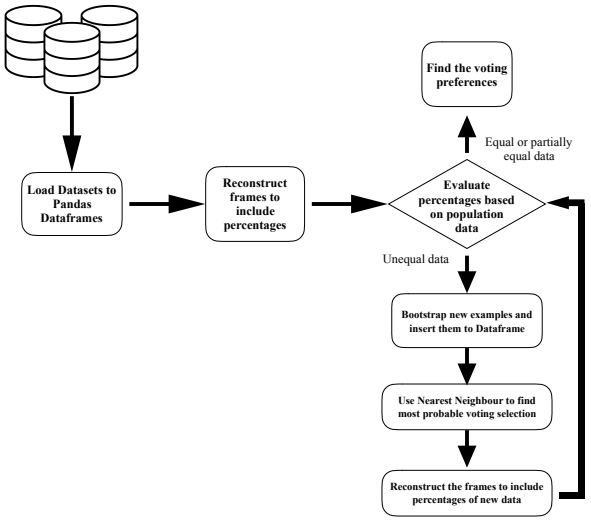

Fig. 2: A simple display of the system's pipeline

The parent iteration that composes all the fields assesses all the flags that were found in each feature. After a specific number of flags is raised the system will stop as it has reached the goal. It can be seen at the code that after each main iteration circle and after considering every field,

a display of the new voting results from the first dataset is shown. This is done since the total process requires a respectable amount of time and therefore sufficient results may be seen before the system has completed the entire task and therefore can be interrupted.

V. EXPERIMENTS

The experiments that were carried out were mainly focused towards the relationships between the six fields previously described with the outcome of the results and also how representative is the new sample created, in comparison with the actual results from the referendum. The most significant improvement in the results was seen during the bootstrapping of both the qualification levels obtained by individuals and also the social index that they have been assigned. It was seen that the data received from YouGov had a larger amount of A and B social grades than what the population data showed in which the biggest concentrations were found in the upper middle class and in the working class as well (coming second). Also the qualification levels were not representative for the total population with many example only been assigned basic qualifications.

```
########## BOOTSTRAPPING NEWSPAPERS ##########
[ 0.024  0.004  0.013  0.263  0.086  0.025  0.177  0.061  0.032  0.007
  0.001  0.069  0.086  0.096  0.057]
[ 0.04   0.017  0.006  0.071  0.05   0.051  0.163  0.107  0.13   0.007
  0.04   0.183  0.067  0.061  0.014]
New number of examples: 17266
2    7006
1    6807
3    2683
4     746
Name: EUREF_Int, dtype: int64

########## BOOTSTRAPPING TV NEWS ##########
[ 0.701  0.     0.16   0.047  0.067  0.024]
[ 0.32   0.226  0.113  0.068  0.17   0.105]
New number of examples: 34533
2    13968
1    13760
3     5347
4     1434
Name: EUREF_Int, dtype: int64


##########
New number of examples: 34533
##########
```

Fig. 3: Partial output from a system run by only using bootstrap. The votes can be seen as the four labels (1:Remain, 2:Leave, 3:Undecided, 4:Would not vote)

The first experiments only used bootstrapping for the original data as it can be observed in Figure 3. However, by solely generating new examples with the voting class to be randomised, the end result will only be a larger representation of the first dataset that will nevertheless have characteristics of the one from the population. Also, the information between features will be lost as the new examples will be built considering independence between the features. For this reason, the next step was to introduce an approach that will allow some information to be preserved in iterations. The Nearest Neighbour method was chosen for the simple implementation that it allows and also the fact that it also includes a weighting factor for each new bootstrapped example based on the similarity with previous examples.

```
########## BOOTSTRAPPING SOCIAL GRADE #########
[ 0.292  0.298  0.203  0.207]
[ 0.228  0.308  0.208  0.257]
New number of examples: 4562
2    1861
1    1699
3     780
4     198
Name: EUREF_Int, dtype: int64
0.410092551785
0.37439400617
0.171881886294
0.0436315557514

########## BOOTSTRAPPING PAST VOTES ##########
[ 0.352  0.317  0.075  0.045  0.005  0.119  0.048  0.001  0.037]
[ 0.369  0.304  0.079  0.047  0.006  0.126  0.04   0.001  0.031]
New number of examples: 5461
2    2234
1    2048
3     922
4     233
Name: EUREF_Int, dtype: int64
0.41088835755
0.376678315247
0.169578811845
0.0428545153577

########## BOOTSTRAPPING EDUCATION QUALIFICATIONS ##########
[ 0.095  0.067  0.166  0.234  0.319  0.118]
[ 0.145  0.156  0.178  0.153  0.293  0.077]
New number of examples: 8164
2    3343
1    3072
3    1379
4     346
Name: EUREF_Int, dtype: int64
0.410687960688
0.377395577396
0.16941031941
0.0425061425061
```

Fig. 4: Second output from a system run with bootstrapping and nearest neighbour for the referendum intention. The votes can be seen as the four labels (1:Remain, 2:Leave, 3:Undecided, 4:Would not vote). Based on these labels the percentages are also calculated for additional proficiency

VI. EVALUATION

As it can be seen by the two runs, the case in which only bootstrap is applied will simply increase the sample size without having any significance in the voting results. In the second case however, by increasing the bias of the method with the introduction of the prediction function for the bootstrap examples, the results show that there is a strong correlation between the socioeconomic data of the population and the actual outcomes. If we use the output shown in Figure 4, but only take to account the Remain and Leave votes, we arrive at a situation of 52.11% for Leave and 47.89% for Remain (which a divergence of approximately 0.02% from the real results).

VII. CONCLUSIONS

The tests up to this point showed promising results for the method that it was used. However, in order to have a more robust image, other data should be used as well as other prediction methods, for new examples. By only considering conditional independence between the features of the examples, we can obtain a good approximation, even if it is incorrect some of the times[16]. Last but not least, possible differences in the results generated can be due to:

- Interaction between the variables of the examples that were not regarded
- Voters that were found to be "undecided" in the questionnaire and therefore a separate predictive system is needed to find the most probable vote intention.
- Criteria that may have been important and they were not bootstrapped

TABLE II: Datasets and their features table (as found in the recorded main iteration circle)

Methodology	Leave/Remain ratio	Error%
Bootstrap	1.015	1.51
Bootstarp & NN	1.108	0.02

APPENDIX

The Github repository for the project can be found by following the link in HTTPS: https://github.com/asterga/ce888assignment.git
or with SSH:
git@github.com:asterga/ce888assignment.git

REFERENCES

[1] S. Herbst, "Numbered voices: How opinion polling has shaped American politics", University of Chicago Press, 1993

[2] P. Lynn and R. Jowell, How Might Opinion Polls be Improved?: The Case for Probability Sampling, Journal of the Royal Statistical Society. Series A (Statistics in Society), vol. 159, no. 1, p. 21, 1996

[3] B. Efron, Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods, Biometrika, vol. 68, no. 3, pp. 589599, 1981

[4] T.J. DiCiccio, and B. Efron, "Bootstrap confidence intervals", Statistical science, pp.189-212, 1996

[5] S.J. Rosenstone, "Forecasting presidential elections", Yale University Press, 1983

[6] P. Gomme, "Iowa electronic markets", Federal Reserve Bank of Cleveland, 2003

[7] A.G. Cuzn, J.S. Armstrong, and R. Jones, "Combining methods to forecast the 2004 presidential election: The Pollyvote", In annual meeting of the Southern Political Science Association, New Orleans, 2005.

[8] J.E. Campbell, "Polls and votes: the trial-heat presidential election forecasting model, certainty, and political campaigns", American Politics Quarterly, 24(4), pp.408-433, 1996

[9] D. Chin, A. Zappone, and J. Zhao, "Analyzing Twitter Sentiment of the 2016 Presidential Candidates", 2016

[10] Election 2015, BBC News. [Online]. Available: http://www.bbc.co.uk/news/election/2015/results

[11] I. M. Foundation, 2011 Census Analysis, Local Area Analysis of Qualifications Across England and Wales, [ARCHIVED CONTENT] UK Government Web Archive The National Archives. [Online]. Available: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-352084

[12] Statistical bulletin:2011 Census: Population Estimates for the United Kingdom, March 2011, Office for National Statistics. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuspopulationestimates\discretionary{-}{}{}fortheunitedkingdom/2012-12-17

[13] Social Grade, National Readership Survey. [Online]. Available: http://www.nrs.co.uk/nrs-print/lifestyle-and-classification-data/social-grade/

[14] Market overview, UK newspaper circulation and readership figures. [Online]. Available: http://www.newsworks.org.uk/Market-Overview

[15] All the latest reports, BARB. [Online]. Available: http://www.barb.co.uk/viewing-data/

[16] A.P. Dawid, "Conditional independence in statistical theory", Journal of the Royal Statistical Society. Series B (Methodological), pp.1-31, 1979