Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

# Analyzing Political Opinions and Prediction of Voting Patterns in the US Election with Data Mining Approaches

Md. Sohel Ahammed[1], Md. Nahid Newaz[2] and Arunavo Dey[3]

[1] Bangladesh University of Business and Technology (BUBT)

*Received: 16 December 2018 Accepted: 5 January 2019 Published: 15 January 2019*

## Abstract

Data is the precious resources. Data contains the useful patterns which provide the crucial information about the prediction of what is going to be happened in the next. In this paper, we aim to identify the political preferences and tendency of the US populations using classification and data mining techniques. To provide the usefulness of proposed model we analyze the electoral data sets in US election obtained from the official website which contains the information about 1984 United States Congressional voting records. This paper shows the classification techniques that can be used to predicting voting patterns in the US House of Representatives and shows the close correspondence between election results and extracted opinion. This paper also shows the political support of the voters and prediction the characteristics of the voter with their political tendency.

***Index terms—***

## 1 Introduction

lection is important because it allows the electorate to decide who's going to make decision for their country for the next couple of years. But this election can be forecasted with a reasonable accuracy. Forecasting election using small polling system is very common approach but this often do not produce reasonable accuracy.

Data mining is a process that examines large preexisting databases in order to generate new information. There are also various works that uses data mining approaches to predict various types of results such as weather forecasting, sports result prediction, future buying decision prediction, etc. But there are very few works that uses data mining approaches to predict voting patterns on election. In this work, we uses data mining approaches to predict voting patterns in USA election. For this study we uses data preprocessing for removing missing value, identifying best attributes and removing duplicate values. We split the dataset into training datasets and test datasets. Then we applied four algorithms Tree J48, Naïve Bayes Classifier, Trees Random Forest and Rules zero or Classifier for predicting voting patterns and also compares the results of those model and finds the best models from those models.

## 2 II.

## 3 Related Works

Gregg R. Murray and Anthony Scime uses data mining approaches to predict individual voting behavior including abstention with the intent of segmenting the electorate in useful and meaningful ways [1]. Gregg R. Murray, Chris Riley, and Anthony Scime, in another study, uses iterative expert data mining to build a likely voter model for presidential election in USA [2]. Bae, Jung-Hwan, Ji-Eun, Song, Min uses Twitter data for predicting trends in South Korea Presidential Election by Text Mining techniques [3]. Tariq Mahmood, TasmiyahIqbal, Farnaz Amin, WaheedaLohanna, Atika Mustafa uses Twitter data to predict 2013 Pakistan Election winner [4].

# 4 III.

Data Preprocessing

# 5 Experimental Methodology

We used 4 algorithms and 8 models (2 models for each algorithm) to predict the voting pattern in the US election. We then analyse and compare the results of those models and finds the best models with most accuracy. The algorithms which are applied for generating models are given below.

i. Trees J48 ii.

Naive From the above table, the best model was identified based on the value of the parameters accuracy, precision, recall, sensitivity, and specificity. The higher the value of accuracy, precision, recall and (sensitivity> specificity), the higher the rank.

# 6 VI.

# 7 Conclusion

Though there are lot of techniques and methods for predicting voting patterns, data mining is the most efficient and effective methods in this fields. In our study, we clearly found that among various data mining algorithms Trees Random Forest performs the best with 98.17% accuracy. In future, we will expand our research in most recent dataset for validating our findings with recent ones. [1] [2]

**1**

E I. Handling with Missing Attributes: In this section, we uses the technique of replacing missing values with mean, median or mode. We uses this approach because it is better approach when the dataset is small and it can prevent data loss. II. Removing Duplicates: We used WEKA tools for removing duplicates from the datasets. We used Remove Duplicates () function in WEKA for removing duplicates. III. Best Attributes Selection: We used Gain Ratio Attribute Eval which evaluates the worth of an attribute by measuring the gain ratio with respect to the class and Ranker which Ranks attributes by their individual evaluations. The top 12 attributes from the whole dataset according to rank from the attributes are presented in Figure 1.



Year 2019
37
Volume XIX
Issue II
Version I ( ) C
Global Journal of Computer Science and Technology

Figure 1: Table 1 :

[1] © 2019 Global Journals

[2] Table 17: Model-8 Precision, Recall, F-measure rate according to Democrat class

**2**

| | | |
|---|---|---|
| iii. | Trees RandomForest | |
| iv. | Rules ZeroOR Classifier | |
| a) Trees J48 | | |
| | We used Model 1 for training dataset and Model | |
| 2 for test dataset evaluation. | | |
| Evaluation of Model 1 Training dataset is given below: | | |
| Bayes classifier | | |
| Correctly Classified Instances | 421 | 96.7816% |
| Incorrectly Classified Instances | 14 | 3.2184% |
| Kappa statistics | 0.9324 | |
| Mean Absolute Error | 0.0582 | |
| Root Mean Squared Error | 0.1706 | |
| Relative Absolute Error | 12.2709% | |
| Root Relative Squared Error | 35.0341% | |
| Total Number of Instances | 435 | |

Figure 2: Table 2 :

**3**

| TP Rate | FP Rate | Precision | Recall | F measures | MCC | ROC Area | PRC area | Class |
|---|---|---|---|---|---|---|---|---|
| 0.966 | 0.030 | 0.981 | 0.966 | 0.974 | 0.938 | 0.975 | 0.973 | democrat |
| Sensitivity & Specificity Calculation for Training Data (Model 1) | | | | | | | | |
| Formula of Sensitivity = TP/ (TP+FN) | | | | | | | | |
| Formula of Specificity = TN/ (TN+FP) | | | | | | | | |
| So Sensitivity = TP Rate = 0.966 & Specificity = 0.030 | | | | | | | | |
| Evaluation of Model 2 test dataset is given below | | | | | | | | |

Figure 3: Table 3 :

**4**

| | | |
|---|---|---|
| Correctly Classified Instances | 105 | 96.3303% |
| Incorrectly Classified Instances | 4 | 3.6697% |
| Kappa statistics | 0.921 | |
| Mean Absolute Error | 0.0619 | |
| Root Mean Squared Error | 0.1894 | |
| Relative Absolute Error | 13.2259% | |
| Root Relative Squared Error | 39.4312% | |
| Total Number of Instances | 109 | |

Figure 4: Table 4 :

**6**

| | | |
|---|---|---|
| Correctly Classified Instances | 395 | 90.8046% |
| Incorrectly Classified Instances | 40 | 9.1954% |
| Kappa statistics | 0.8094 | |
| Mean Absolute Error | 0.0965 | |
| Root Mean Squared Error | 0.2921 | |
| Relative Absolute Error | 20.34% | |
| Root Relative Squared Error | 59.9863% | |
| Total Number of Instances | 435 | |

Figure 5: Table 6 :

**7**

| TP Rate | FP Rate | Precision | Recall | F measures | MCC | Rock Area | PRC area | Cla |
|---|---|---|---|---|---|---|---|---|
| 0.895 | 0.071 | 0.952 | 0.895 | 0.923 | 0.812 | 0.972 | 0.983 | dem |

Sensitivity & Specificity Calculation for Training Data (Model 3)
So Sensitivity = TP Rate = 0.895 & Specificity = 0.071
Evaluation of Model 4 test dataset is given below

Figure 6: Table 7 :

**8**

| | | |
|---|---|---|
| Correctly Classified Instances | 99 | 90.8257% |
| Incorrectly Classified Instances | 10 | 9.1743% |
| Kappa statistics | 0.8069 | |
| Mean Absolute Error | 0.0978 | |
| Root Mean Squared Error | 0.2934 | |
| Relative Absolute Error | 20.9083% | |
| Root Relative Squared Error | 61.0861% | |
| Total Number of Instances | 109 | |

Figure 7: Table 8 :

**9**

| TP Rate | FP Rate | Precision | Recall | F measures | MCC | Rock Area | PRC area | Class |
|---|---|---|---|---|---|---|---|---|
| 0.886 | 0.051 | 0.969 | 0.886 | 0.925 | 0.812 | 0.969 | 0.984 | democrat |

Sensitivity & Specificity Calculation for Model 4
Sensitivity = TP Rate = 0.886 & Specificity = 0.051
c) Trees Random Forest
Evaluation on Training Data set: Trees Random Forest algorithm

Figure 8: Table 9 :

**10**

Analyzing Political Opinions and Prediction of Voting Patterns in the US Election with Data Mining Approaches

Year 2 019
39
Volume XIX Issue II Version I
( ) C
Global Journal of Computer Science and Technology

| | | |
|---|---|---|
| Correctly Classified Instances | 427 | 98.1609% |
| Incorrectly Classified Instances | 8 | 1.8391% |
| Kappa statistics | 0.9613 | |
| Mean Absolute Error | 0.0376 | |
| Root Mean Squared Error | 0.1222 | |
| Relative Absolute Error | 7.9365% | |
| Root Relative Squared Error | 25.0915% | |
| Total Number of Instances | 435 | |



Figure 9: Table 10 :

**5**

| TP Rate | FP Rate | Precision | Recall | F measures | MCC | Rock Area | PRC area | Class |
|---|---|---|---|---|---|---|---|---|
| 0.981 | 0.018 | 0.989 | 0.981 | 0.985 | 0.961 | 0.998 | 0.999 | democrat |

Sensitivity & Specificity Calculation for Training Data (Model 5)
So Sensitivity = TP Rate = 0.981& Specificity = 0.018
Evaluation of Model 6 test dataset is given below

Figure 10: Table 5 :

**12**

| | | |
|---|---|---|
| Correctly Classified Instances | 106 | 97.2477% |
| Incorrectly Classified Instances | 03 | 2.7523% |
| Kappa statistics | 0.9404 | |
| Mean Absolute Error | 0.0432 | |
| Root Mean Squared Error | 0.1508 | |
| Relative Absolute Error | 9.2437% | |
| Root Relative Squared Error | 31.408% | |
| Total Number of Instances | 109 | |

Figure 11: Table 12 :

**13**

| TP Rate | FP Rate | Precision | Recall | F measures | MCC | Rock Area | PRC area | Cla |
|---|---|---|---|---|---|---|---|---|
| 0.971 | 0.026 | 0.986 | 0.971 | 0.978 | 0.941 | 0.996 | 0.997 | der |

Sensitivity & Specificity Calculation for Model 6
Sensitivity = TP Rate = 0.971 & Specificity = 0.026
d) Rules ZeroOR Classifier
Evaluation on Training Data set: Rules ZeroOR Classifier algorithm

Figure 12: Table 13 :

**14**

| | | |
|---|---|---|
| Correctly Classified Instances | 267 | 61.3793% |
| Incorrectly Classified Instances | 168 | 38.6207% |
| Kappa statistics | 0 | |
| Mean Absolute Error | 0.4742 | |
| Root Mean Squared Error | 0.4869 | |
| Relative Absolute Error | 100% | |
| Root Relative Squared Error | 100% | |
| Total Number of Instances | 435 | |

Figure 13: Table 14 :

**15**

| TP Rate | FP Rate | Precision | Recall | F measures | MCC | Rock Area | PRC area | Class |
|---|---|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 0.614 | 1.0 | 0.761 | - | 0.500 | 0.614 | democrat |

Sensitivity & Specificity Calculation for Training Data (Model 7)
So Sensitivity = TP Rate = 1.0 & Specificity = 1.0
Evaluation of Model 8 test dataset is given below

Figure 14: Table 15 :

**16**

Analyzing Political Opinions and Prediction of Voting Patterns in the US Election with Data Mining Approaches

Year 2 019
40
Volume XIX Issue II Version I
)
( C
Global Journal of Computer Science and Technology

| | | |
|---|---|---|
| Correctly Classified Instances | 70 | 64.2202% |
| Incorrectly Classified Instances | 39 | 35.7798% |
| Kappa statistics | 0 | |
| Mean Absolute Error | 0.4678 | |
| Root Mean Squared Error | 0.4802 | |
| Relative Absolute Error | 100% | |
| Root Relative Squared Error | 100% | |
| Total Number of Instances | 109 | |



Figure 15: Table 16 :

**11**

| TP Rate | FP Rate | Precision | Recall | F measures | MCR | Rock Area | PRC area | Clas |
|---|---|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 0.642 | 1.0 | 0.782 | - | 0.500 | 0.642 | dem |

Sensitivity & Specificity Calculation for Model 8
Sensitivity = TP Rate = 1.0& Specificity = 1.0
V. Revaluation of the Best, Second Best and Third Best Model

Figure 16: Table 11 :

**18**

Figure 17: Table 18 :

| Model | Accuracy | precision | recall | 0.981 | 0.966 | 0.985 | 0.957 | sensitivity | specificity | 0.966 | 0.030 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 96.7816% | | | | | | | | | | | |
| Model 2 | 96.3303% | | | | | | | | | | | |
| Model 3 | 90.8046% | 0.952 | | 0.895 | | | | 0.895 | | 0.071 | | |
| Model 4 | 90.8257% | 0.969 | 0.989 | 0.886 | | | | 0.886 | 0.981 | 0.051 | | |
| Model 5 | 98.1609% | 0.986 | 0.614 | 0.985 | | | | 0.971 | 1.00 | 0.018 | | |
| Model 6 | 97.2477% | 0.642 | | 0.978 | | | | 1.00 | | 0.026 | | |
| Model 7 | 61.3793% | | | 1.00 | | | | | | 1.00 | | |
| Model 8 | 64.2202% | | | 1.00 | | | | | | 1.00 | | |

Figure 18:

[Bae and Song ()] ‘Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques’. Jung-Hwan Bae , Ji-Eun Song , Min . *Journal of intelligence and Information Systems* 2013. 19 (3) .

[Gregg et al. ()] ‘Micro targeting and Electorate Segmentation: Data Mining the American National Election Studies’. R Gregg , Anthony Murray , Scime . *Journal of political marketing* 2010. 9 (3) .

[Mahmood et al.] *Mining Twitter*, Tariq Mahmood , Farnaz Tasmiyahiqbal , Amin , Atika Waheedalohanna , Mustafa .

[Murray et al. ()] ‘Pre-Election Polling: Identifying Likely Voters Using Iterative Expert Data Mining’. Greg R Murray , Chris Riley , Anthony Scime . *Public opinion Quarterly* 2009. 73 (1) .