Journal of Universal Computer Science, vol. 23, no. 3 (2017), 304-326
submitted: 15/9/16, accepted: 1/3/17, appeared: 28/3/17 © J.UCS

Electoral Preferences Prediction of the YouGov Social Network Users Based on Computational Intelligence Algorithms

Sonia Ortiz-Ángeles
(Instituto Politécnico Nacional, CIDETEC-IPN, Ciudad de México, México
lizde00@yahoo.com)

Yenny Villuendas-Rey
(Instituto Politécnico Nacional, CIDETEC-IPN, Ciudad de México, México
yenny.villuendas@gmail.com)

Itzamá López-Yáñez
(Instituto Politécnico Nacional, CIDETEC-IPN, Ciudad de México, México
ilopezy@ipn.mx)

Oscar Camacho-Nieto
(Instituto Politécnico Nacional, CIDETEC-IPN, Ciudad de México, México
oscarc@cic.ipn.mx)

Cornelio Yáñez-Márquez
(Instituto Politécnico Nacional, CIC-IPN, Ciudad de México, México
cyanez@cic.ipn.mx)

Abstract: The contemporary world has witnessed technological advances, such as Online Social Networks (OSN), whose influence in almost every action of the human being is remarkable. Among the human activities most significantly impacted by OSNs are: entertainment, human relationships, education, and political activities, including those related to electoral campaigns and electoral preferences prediction. The research contribution of the current paper regards the usefulness of OSNs users generated data to predict the political context. More specifically, 25 Computational Intelligence (CI) algorithms are used to predict voting intentions on the United States primary presidential elections for 2016, taking as input the data sets generated by 1200 users of the YouGov OSN, as well as the answers they gave to an online study run by the American National Election Studies (ANES). The application of the 25 supervised classification algorithms is done over the Waikato Environment for Knowledge Analysis (WEKA), using a stratified 5-fold cross validation scheme. Also, the experimental results obtained were validated in order to identify significant differences in performance by mean of a non-parametric statistical test (the Friedman test), and a post-hoc test (the Holm test). The hypothesis testing analysis of the experimental results indicates that predicting voting intentions in favour of a democrat or republican candidate is simpler than predicting the particular candidate, given that the prediction performances for a democrat or republican candidate (best performances of 80% and 78%, respectively) are better than those given when predicting a specific candidate (70% for democrat candidates and 56% for republican candidates).

Keywords: electoral preferences, prediction, online social networks, computational intelligence
Categories: H.3.5, I.5.1, I.5.4, J.4, K.4.2, L.6.1

1 Introduction

In recent years, an increasing and accelerating wave of technological advances has been introduced into practically all human activities, influencing almost every action of individuals and organizations, and generating changes to common practices and attitudes. Modern digital computers, smart phones, and tablets are powerful devices able to process and move the user information in a friendly manner, with applications supported by the advances in communications technologies [Niebert et al., 07].

Like [Webster and Murphy, 08] novel technologies are constantly emerging, and examples of these are: virtual worlds, mobile devices, wireless ad hoc networks, open source software developments, management systems, and OSN; which have a great influence on the activities of the contemporary human being. Considering that millions of OSN users produce huge amounts of information through their daily interactions, it becomes clear the large number of activities on which OSN impact in the current world [Jung and Kazienko, 12]. Among this wide variety of activities, some of the most significantly outstanding are entertainment and human relationships [Montag and Reuter, 15], education [Bicen and Uzunboylu, 13], and political activities, more specifically those related to electoral campaigns, vote promotion and electoral preferences prediction [Espinosa-Oviedo et al., 16].

A community seen as an OSN consists of members reciprocally interacting with each other. These interactions produce cohesion between members while being inaccessible to outsiders. The OSN do not limit membership to people within the same location neither require everyone to be connected at the same time. Members in an OSN can also provide other resources to each other: information, feedback, advice, job opportunities, and news, among others [Smailovic and Podobnik, 16].

Even though Facebook and Twitter are undoubtedly the most well-known and popular OSNs, actually an amazing variety of OSNs have arisen to cater to every use and preference, where every social networking site is unique in its application or characteristics. Ello, Medium, Reddit, Poolwo, Livejournal, Pinterest, Quora, Google Plus, and StumbleUpon, are but a meagre sample [Zidan, 16].

In the current paper, several supervised learning algorithms are used to predict electoral preferences, taking as input data sets generated by the YouGov OSN users [https://today.yougov.com].

The main contribution of this paper is to provide a deep analysis of the performance of supervised classification algorithms in the prediction of electoral preferences of YouGov users for the 2016 US presidential primary. In addition, the paper provides insights about the voting intentions of users, as well as the relevant attributes needed to achieve an accurate prediction.

The rest of this paper is organized as follows: section 2 is dedicated to presenting the YouGov OSN and describing the data sets used in this research. The third section describes the measures of performance used to evaluate the Computation Intelligence algorithms compared in the experimental study. These algorithms and the software platform employed to execute them are discussed in section 4. Meanwhile, section 5 is dedicated to presenting the statistical analysis tools applied to the experimental results, and section 6 contains the experimental results and their discussion; leaving conclusions and future work for section 7, and finally the references are included.

2 Related works

The electoral question, as well as the majority of social phenomena, is a difficult subject to model from a statistic-mathematical point of view, since the relations between variables in social sciences rarely respond to a linear behavior pattern, or to a predetermined mathematical model [Little and Rubin, 89]. In addition, the multidimensionality of social phenomena forces more attention to the interaction and mutual influence of the intervening variables [Bohrnstedt and Knoke, 82].

Regarding the electoral question, several studies have been carried out in two areas: to predict the winner in a particular election [Mattes and Milazzo, 14; Banai et al., 16] and to understand the electoral behavior [Bartel,s 00; Fisher et al., 15]. In the latter case, the data are usually obtained through surveys. Several countries have institutions dedicated to this type of studies, such as United States, France, the United Kingdom and Sweden.

In the United States, for example, the National Science Foundation [https://www.nsf.gov/] funds studies of the presidential and legislative elections held by the Center for Political Studies at the University of Michigan. Also, since 1948, the American National Election Studies (ANES) [http://www.electionstudies.org/] has conducted surveys, in most national election years; in addition, in recent years ANES had used information technologies to this end, by using on-line panel surveys, as well as exploiting Social Network sites and OSN [Thelwall, 09].

In France, the Institute of Political Sciences of Paris and the National Foundation of Political Sciences [http://www.sciencespo.fr/] have carried out important surveys of this type. In Great Britain, the Social Science Research Council provides funds for conducting academic surveys. Also, in Sweden, the government sponsors two types of academic surveys: general and annual surveys conducted by the National Bureau of Statistics; and electoral studies at the national level [http://www.scb.se/en/finding-statistics/statistics-by-subject-area/democracy/] which are now carried out jointly by the National Directorate of Statistics.

Following a different perspective, [Banai et al., 16] use voice feature of the candidates to predict which of them will be more likely voted by the electorate. They findings suggest that “candidates with lower-pitched voices had greater likelihood of winning the election if they had higher pitch variability.”

On the other hand, OSN have recently been considered of predictive value, and their use for political discourse is becoming a more common practice, especially in times of elections [You et al., 15; Cameron et al., 16]. It could be argued that one of the most interesting aspects of this trend is the possibility of obtaining a "pulse" of public opinion almost in real time and, therefore, has attracted the interest of various researchers as well as news organizations. It is for this reason that the idea has been consolidated that a prediction of electoral results could be obtained from the data generated in OSN.

Several researches had addressed the issue of predicting election results based on twitter comments, in countries such as German [Tumasjan et al., 10], the Netherlands [Sang et al., 12], Singapore [Skoric et al., 12], and Greece [Tsakalidis et al., 15]. However, the capability of tweets for this particular task is not conclusive, since there are some success papers, as well as some negative-results papers about it. In addition, [Gayo-Avello et al., 11] have pointed out some of the limitations of the predictive power of social media data, due to several difficulties around political writing: “it is plagued with humour, double intender, and sarcasm” [Gayo-Avello, 12].

3 The YouGov OSN and the datasets

The data sets used in this research were compiled from online information provided by users of the YouGov social network [https://today.yougov.com], which according to its own website “is a global online community, where millions of people and thousands of political, cultural and commercial organizations engage in a continuous conversation about their beliefs, behaviours and brands”. Among their services is the YouGov Profile [https://today.yougov.com/find-solutions/profiles/], which is described as “a new tool for media planning, segmentation and forecasting”.

YouGov profiles include information about attitudes and opinions of the users affiliated to this OSN. These profiles also offer information regarding several topics of interest to the present work, such as vote intention, political leanings, and personal beliefs.

In particular, 20 variables of the YouGov Profile of 1200 users were used for this work, as well as the answers provided by these users to an online questionnaire [http://electionstudies.org/studypages/anes_pilot_2016/anes_pilot_2016.htm] applied by the American National Election Studies (ANES) between February the 22 and 28, 2016. The goal of this questionnaire is to obtain information about the voting intentions of the OSN users in the then upcoming primary elections for the US President in 2016.

The online ANES study is composed by 213 questions (variables) about the opinions of the YouGov OSN users regarding economical, racial, violence, world trends, and other political topics, as well as the voting intentions of the users for the primary presidential elections. Also, the study reports include the answering time for each question by user, besides the browser and operating system used to fill out the online questionnaire.

Three different scenarios were considered for the data analysis, in terms of the selection of the variables to be used:

1. Use only data taken from the YouGov user profile.
2. Use only data taken from the online questionnaire answered by the YouGov OSN user.
3. Use data taken from the YouGov profile and the ANES questionnaire.

In terms of the voting intentions prediction, four scenarios were taken into account:

1. Predict the candidate for which the user will vote, from a list of possible democrat candidates (variable *demcand* of the online study). In this case, the available classes to predict are five: “Hillary Clinton, Martin O’Malley, Bernie Sanders, Another Democratic Candidate, None”.
2. Predict the candidate for which the user will vote, from a list of possible republican candidates (variable *repcand* of the online questionnaire). In this case, the available classes to predict are 11: “Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, John Kasich, Rand Paul, Marco Rubio, Donald Trump, Another Republican Candidate, None”.

3. Predict whether the user will vote for a democrat candidate or not (variable *demcand* of the online study). In this case there are two classes to predict, considering all democrat candidates as one class, and the "None" vote as the contrary class.
4. Predict whether the user will vote for a republican candidate or not (variable *repcand* of the online questionnaire). In this case there are two classes to predict, considering all republican candidates as one class, and the "None" vote as the contrary class

Thus, there are a total of 12 data sets, of which six correspond to a two classes problem, while the other six correspond to a multiple class problem. Table 1 presents the description of the 12 different datasets. Notice that all datasets include missing values.

No.	Datasets	Attributes		Imbalance analysis			Classes
		Num.	Cat.	Majority class	Minority class	IR	
1.	demcand_profile	1	16	366	63	5.81	5
2.	demcand_quest	29	183	366	63	5.81	5
3.	demcand_quest_profile	30	199	366	63	5.81	5
4.	repcand_profile	1	16	355	22	16.14	11
5.	repcand_quest	29	183	355	22	16.14	11
6.	repcand_quest_profile	30	199	355	22	16.14	11
7.	dem_profile	1	16	862	337	2.56	2
8.	dem_quest	29	183	862	337	2.56	2
9.	dem_quest_profile	30	199	862	337	2.56	2
10.	rep_profile	1	16	845	355	2.38	2
11.	rep_quest	29	183	845	355	2.38	2
12.	rep_quest_profile	30	199	845	355	2.38	2

Table 1: Description of the data sets used.

As can be seen in table 1, the data sets drawn from the online study answered by YouGov OSN users contain mixed attributes (both numerical and categorical), as well as missing values.

It is also evident that the number of users who would vote for different candidates is quite different (e.g. only 22 indicate a voting intention in favor of republican candidate Carly Fiorina), which makes the imbalance ratio (IR) between classes greater than 2 for every case, which in turn indicates that the 12 data sets are imbalanced.

Considering the imbalance ratio, the data for the democrat candidates (datasets 1-3) is distributed with a majority of 366 votes for a candidate, and a minority of 63 votes for a candidate. In a similar manner, the data for the republican candidates (datasets 4-6) is distributed with a majority of 355 votes for a candidate, and a minority of 22 votes for a candidate.

However, the data for the democrat vs. republican and republican vs. democrat (datasets 7-9 and 10-11, respectively) the data is more equally distributed, having a relation of majority-minority votes of (862 votes – 337 votes) and (845 votes – 355 votes), respectively.

Regarding the attributes, the variables *race_other*, *employ_t*, and *faminc2* were eliminated from the YouGov profile, given that none of the user offered information on such attributes.

On the other hand, in the data set focused on analyzing the voting preference for democrat candidates (data sets 1-3 and 7-9), one user who did not answer the question was eliminated, leaving such data sets with a total of 1199 users.

In order to illustrate the previous discussion, let us consider the following example, which is a summary of the attribute values of user 1118 from the dataset 1.

In this case (figure 1, dataset 1), there are 5 classes, of which the user indicated to have a voting preference for candidate Bernie Sanders. This same user would appear in the dataset 7 (two classes problem) as having indicated a voting preference for a democrat candidate.

ATTR. NUM.	1	2	3	…	15	16	17	18
ATTR. NAME	*birthyr*	*gender*	*race*	…	*pew_churatd*	*religpew*	*religpew_t*	**Class**
VALUE	1962	Female	White	…	Once a week	Protestant	Baptist	**Bernie Sanders**

Figure 1: Example of the data corresponding to one user in dataset 1.

Supervised classification algorithms used for the prediction of voting intention will take this information, learning the attribute values of a user and associating this with the corresponding class (i.e. voting intention) during the training phase of the algorithm. During the classification phase, a similar vector of attribute values (corresponding to an unknown user) will be presented to the algorithm without the class information. If the algorithm outputs a voting intention equal to the one indicated by the unknown user, the result is considered correct for this user. Instead, if the class offered by the algorithm is different from the one chosen by the user, the result is considered incorrect. This process is shown in the figure 2.

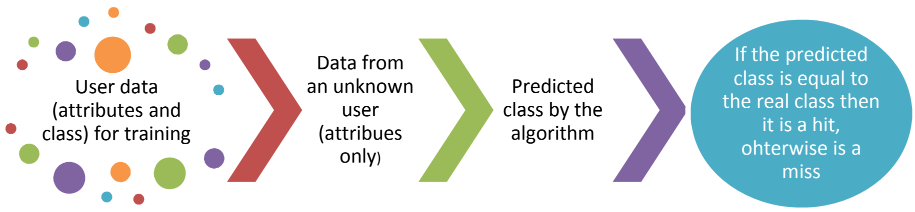

Figure 2: Process of intelligent prediction of voting intention.

Some standards that should be follow to predict elections from social media data are proposed by [Metaxas et al., 11], and are summary as “The prediction theory should be an algorithm with carefully predetermined parameters, the data analysis should be aware of the difference between social media data and natural phenomena data, and it should contain some explanation on why it works.”

Considering the above-mentioned standards, we satisfied the first standard, due to we test some well-known and well defined algorithms (explained in section 5). We also take into account the differences about social media data and natural phenomena data, due to our datasets come from profile data and questionnaire data, collected carefully by YouGov social network and ANES, respectively. In addition, most of the supervised classification algorithms used in this research (section 5) provide an adequate explanation on why they work, while a few (such as neural networks) perform as black boxes.

4 Supervised classification algorithms in WEKA

According to [Hall et al., 09], WEKA is a collection of machine learning algorithms for data mining tasks, including data pre-processing, classification, regression, clustering, association rules, and visualization.

Family	Algorithm (WEKA)	Description (available in WEKA)	Cites
Bayes based	BayesNet	Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier.	[Witten and Frank, 05]
	NaiveBayes	Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data.	[John and Langley, 95]
Logistic based	Logistic	Class for building and using a multinomial logistic regression model with a ridge estimator.	[Le Cessie and van Houwelingen, 92]
Neural Networks	MLP	MultilayerPerceptron. A Neural Network classifier that uses backpropagation to classify instances.	[Witten and Frank, 05]
	RBFNetwork	Implements a normalized Gaussian radial basis function network. It uses the k-means clustering algorithm to provide the basis functions and learns a logistic regression on top of that.	[Witten and Frank, 05]
Support Vector Machines	SMO	Implements John Platt's sequential minimal optimization algorithm for training a Support Vector classifier.	[Platt, 99; Keerthi et al., 01; Hastie and Tibshirani, 98]
Lazy learners	1NN	Nearest Neighbor classifier. Uses Euclidean distance.	[Aha and Kibler, 91]
	3NN	K Nearest Neighbors classifier. Uses Euclidean distance.	[Aha and Kibler ,91]
	Kstar	K* is an instance-based classifier. It differs from other instance-based learners in that it uses an entropy-based distance function.	[Cleary and Trigg, 95]
	LWL	Locally weighted learning. Uses an instance-based algorithm to assign instance weights. Uses DecisionStump and a linear Nearest Neighbor search.	[Frank et al., 03, Atkeson et al., 97]
Others	HyperPipes	Class implementing a HyperPipe classifier. For each category a HyperPipe is constructed (essentially records the attribute bounds observed for each	[Witten and Frank, 05]

		category).	
	VFI	Classification by voting feature intervals. Intervals are constructed around each class for each attribute (basically discretization).	[Demiroz and Guvenir, 97]
Rule based	JRip	Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.	[Cohen, 95]
	Nnge	Nearest-neighbor-like algorithm using non-nested generalized exemplars (which are hyperrectangles that can be viewed as if-then rules).	[Brent, 95, Sylvain, 02]
	OneR	Class for building and using a 1R classifier; in other words, uses the minimum-error attribute for prediction, discretizing numeric attributes.	[Holte, 93]
	ZeroR	Class for building and using a 0-R classifier. Predicts the mode (for a nominal class).	[Witten and Frank, 05]
Decision Trees	BFTree	Class for building a best-first decision tree classifier. This class uses binary split for all attributes. For missing values, the method of “fractional instances” is used.	[Shi, 07, Friedman et al., 00]
	DecisionStump	Class for building and using a decision stump. Does classification (based on entropy). Missing is treated as a separate value.	[Witten and Frank, 05]
	J48	Class for generating a pruned or unpruned C4.5 decision tree.	[Quinlan, 93]
	LADTree	Class for generating a multi-class alternating decision tree using the LogitBoost strategy.	[Holmes et al., 01]
	LMT	Classifier for building Logistic Model Trees, which are classification trees with logistic regression functions at the leaves.	[Landwehr et al., 05; Sumner et al., 05]
	NBTree	Class for generating a decision tree with naive Bayes classifiers at the leaves.	[Kohavi, 96]
	RandomTree	Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning.	[Witten and Frank, 05]
	REPTree	Fast decision tree learner. Builds a decision tree using information gain/variance and prunes it using reduced-error pruning.	[Witten and Frank, 05]
	SimpleCART	Class implementing minimal cost-complexity pruning. When dealing with missing values, uses "fractional instances" method.	[Breiman et al., 84]

Table 2: Description of the 25 classification algorithms used in experiments.

This open source tool was developed by a team of researchers at the University of Waikato, in New Zealand, and led by [Witten and Frank, 05]. As an acronym, WEKA stands for “Waikato Environment for Knowledge Analysis”. This software is written in Java, which gives it very good portability, enabling it to run on the three mainstream desktop operating systems: Windows, Linux, and Mac. Weka is distributed under the GNU license, and may be freely downloaded from the following URL: http://www.cs.waikato.ac.nz/ml/weka/

In terms of the algorithms used for the experimental comparison, 25 of the classification algorithms available on WEKA were executed. Table 2 shows a brief description of these algorithms.

5 Performance measures and statistical analysis

Given the class imbalance shown by all data sets, a stratified 5-fold cross validation sampling was chosen, since this technique is considered adequate for handling imbalanced data sets [López et al., 13; Sáez et al., 15; Vluymans et al., 16]. Stratified cross validation consists of partitioning the whole data set into k folds in such manner that each class is equally represented in each fold. Thus, one fold is used as a test set while the rest is used as the training set. This process is repeated k times, exchanging in each instance which folds are used for training and testing. In the case of the current work, 20% of the dataset was used for testing and 80% was used for training in each iteration, since k = 5.

When imbalanced data sets are used for the task of classification, the usual performance measures —such as the rate of correctly classified instances— become inappropriate [Fernández et al., 13]. This is due to the bias that such measures have towards the majority class, since they do not differentiate between the number of correctly classified instances form different classes, which in turn may yield to misleading conclusions. For evaluating the performance over imbalanced data sets with multiple classes, the use of minimum sensitivity [Fernández-Navarro et al., 11] and the average sensitivity per class [Fernández et al., 13] have been proposed.

In a two classes problem, sensitivity (also known as recall or true positive rate TPR) considers the total of positive instances correctly classified, relative to the total of instances of the positive class, considering True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

$$Sensitivity = TPR = Recall = \frac{TP}{TP + FN} \tag{1}$$

However, in a problem with *k* classes the sensitivity takes into account the total of correctly classified instances from class *i*, relative to the total of instances of the *i*-th class. Thus, the sensitivity for class *i* estimates the probability of correctly classifying an instance from class *i*. For the computation of such sensitivity, let n_i be the number of correctly classified instances (in a confusion matrix of *k* classes), and let t_i be the total of instances belonging to class *i*. Then the sensitivity (also recall or true positive rate) of class *i*, denoted by S_i, is computed as follows:

$$S_i = Recall_i = TPR_i = \frac{n_i}{t_i} \tag{2}$$

Thus, the minimum sensitivity is given by [Fernández-Navarro et al., 11]:

$$MS = \min_{i=1..k}\{S_i\} \tag{3}$$

Even though minimum sensitivity allows handling multiple classes, it only considers the lesser of the correctly classified rates among the classes. This is the reason why in this research, a measurement of performance giving the same weight to each of the classes, independently from the number of samples each has, was chosen. The performance measure used here is the average sensitivity per class [Fernández et al., 13] which is defined as:

$$Avg\_S = \frac{1}{k}\sum_{i=1}^{k} S_i \qquad (4)$$

where *k* is the number of classes and S_i is the sensitivity for the *i*-th class. This performance measure enables us to evaluate the global performance of classification algorithms over all the classes in the problem, not only over the minority class. The use of the average sensitivity per class allows taking into consideration all the classes, without bias towards any particular one. Figure 3 shows an example of how the minimum sensitivity and the average sensitivity are computed, with *k* = 3 classes.

		Predicted class		
		A	*B*	*C*
Real class	*A*	10	0	0
	B	3	3	4
	C	1	2	7

$$S_A = \frac{10}{10} = 1,\ S_B = \frac{3}{10} = 0.3,\ S_C = \frac{7}{10} = 0.7$$

a) $Avg\_S = \frac{1+0.3+0.7}{3} = \frac{2}{3}$ b) $MS = min\{1, 0.3, 0.7\} = 0.3$

Figure 3: Example of computation of performance measures: (a) average sensitivity, and (b) minimum sensitivity, given a confusion matrix for three classes.

In order to determine which classification algorithms obtained the better experimental results while predicting voting intentions, hypothesis testing was used. Statistical hypothesis tests evaluate whether there is a significant difference in the performance given by different classification algorithms, in regards to their prediction of the voting intentions. Considering the works of [Demšar, 06; Garcia and Herrera, 08; Garcia et al., 10], non-parametric tests were chosen for the current research. In particular, the Friedman test was selected since it is widely recommended for this kind of studies.

The Friedman test (more specifically the Freidman two-way analysis of variance by ranks) is a statistical non-parametric test developed by Friedman [Friedman, 37; Friedman, 40], which turns out to be the non-parametric equivalent to the two-way ANOVA analysis. The Friedman test consists of ordering the samples, replacing them by their respective ranks: the best result corresponds to rank 1, the second best to rank 2, and so on. When ordering, the existence of identical samples is taken into account, in which case they are assigned an averaged rank.

If the null hypothesis of performance equality is rejected by the Friedman test, the application of a post-hoc test is needed to determine between which algorithms there are differences [Demšar, 06; Garcia et al., 10]. The post-hoc tests consider the z statistic to compare two algorithms. The z value is used to find the corresponding probability value p and compare it to the significance value α. In statistical hypothesis testing, the p value represents the probability of obtaining a result as extreme as the one already observed, assuming the null hypothesis is true. The lesser the p value, the more evidence present against the veracity of the null hypothesis. If the p value is less than the significance level α, the null hypothesis is rejected and it is accepted that significant differences between the two classifier performances exist.

Among the different post-hoc tests recommended for classification algorithms performance analysis over multiple data sets [Demšar, 06; Garcia and Herrera, 08; Garcia et al., 10] we find the Holm test [Holm, 79]. This test uses a descending (step-down) procedure to adjust the significance value α. For this, the p values are ordered ascendingly (i.e. from the most significant to the least significant). If $p_1 < \frac{\alpha}{l-1}$, the null hypothesis is rejected and the test continues the comparison with the next p value, considering whether $p_2 < \frac{\alpha}{l-2}$. The Holm test continues this process until one of the hypothesis cannot be rejected, given that $p_i \geq \frac{\alpha}{l-i}$. At this point, the remaining hypotheses are also not rejected.

There are several automated tools for the computation of the Friedman test, as well as the post-hoc tests. In this work, the KEEL software was used [Alcalá-Fdez et al., 09; Alcalá-Fdez et al., 11].

6 Experimental Results and Discussion

This section presents the experimental results obtained in predicting voting intentions of users of the YouGov OSN [https://today.yougov.com], in the primary presidential elections in 2016 at the United States, based on the data drawn with the ANES online study [http://electionstudies.org/studypages/anes_pilot_2016/anes_pilot_2016.htm]. For this, the performance behaviors of 25 classification algorithms available on the WEKA platform were analyzed [Hall et al., 09]. Figure 4 illustrates the schematics of the experiment design.

6.1 Summary of major findings

The huge experimental comparison carried out, and the subsequent statistical analysis, allow us to understand in a deeper way the performance of supervised classification algorithms in the prediction of electoral preferences of YouGov users for the 2016 US presidential primary. As a contribution, we found that according to the statistical tests, BFTree and CART algorithms are the best for predicting voting intentions for a single candidate and for democrat/republican vote, respectively.

In addition, we found that predicting voting intentions for a democrat candidate is easier than predicting voting intentions for a republican candidate (maximum average sensitivity of 70% and 56%, respectively).

The discussion of results provides insights about the voting intentions of users with respect to democrat and republican candidates, as well as the relevant attributes needed to achieve an accurate prediction. Such results support the assertion that the YouGov user profile includes attributes that differentiate the voting intention for a democrat candidate or not. However, these same attributes are not good enough by themselves to predict voting intentions for particular candidates (regardless of party), nor for predicting a voting preference for a republican candidate or not.

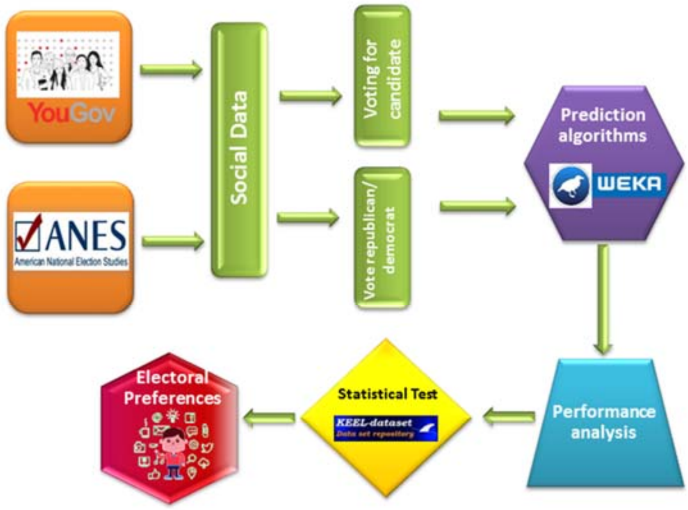

Figure 4: Schematic of the experimental design

The experimental analysis is divided in three parts. In the first part, the prediction of voting preference for a particular candidate is considered (data sets 1-6 from table 1), the second part involves predicting whether the YouGov user intends to vote for a democrat or a republican candidate (data set 7-12 in table 1), while the third part analyzes the attributes needed for accurately predicting voting intentions.

6.2 Results on the prediction of voting intentions for a particular candidate

Notice that the problem of predicting voting intentions for a specific candidate is highly complex, since all candidates to predict share many traits in common, given that all candidates considered in each experiment belong to the same party.

Figure 5 shows the results obtained by the analyzed classification algorithms, on the prediction of voting intention for one specific candidate. Predictions of voting intentions for a democrat candidate have a maximum average sensitivity of 70%. These results are quite relevant, since this is a complex problem. On the other hand, the prediction results for a particular republican candidate are soberer, reaching a maximum average sensitivity of 56%.

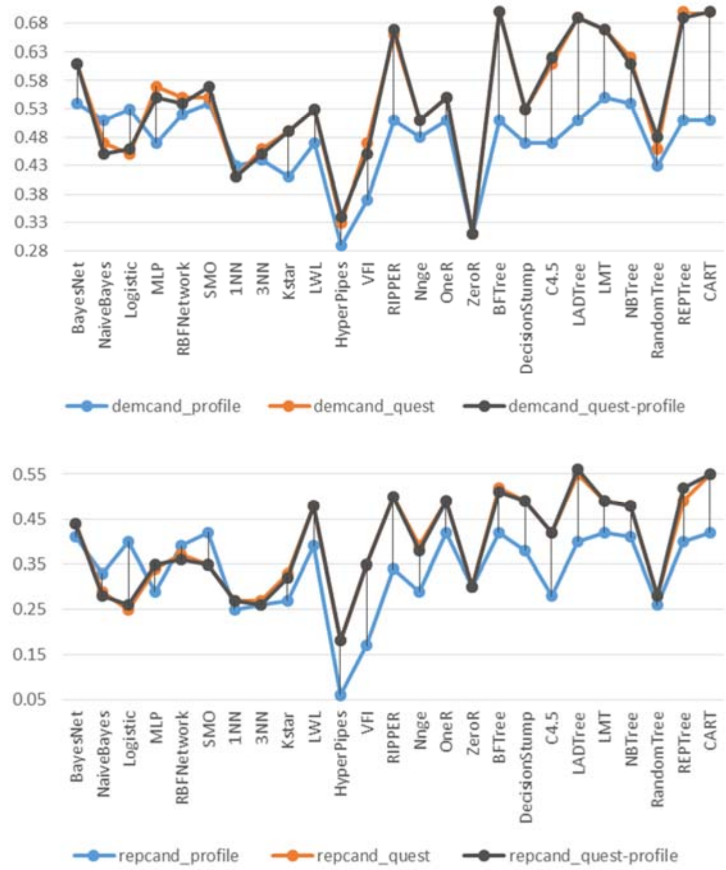

Figure 5: Average sensitivity per class of voting intentions for a particular candidate: democrat (above) and republican (below).

In order to find out which of the algorithms under comparison is more appropriate for the correct prediction of voting intentions, the Friedman test [Friedman, 37; Friedman, 40] was applied, giving a value of $p = 7.58\text{E}^{-11}$, which is largely below the established significance level of $\alpha = 0.05$ for a 95% confidence. The algorithms rankings according to the Friedman test are shown in table 3, where the best classifier for this task is clearly the BFTree [Shi, 07; Friedman et al., 00].

No.	Ranking	Algorithm	No.	Ranking	Algorithm	No.	Ranking	Algorithm
1	**3.000**	**BFTree**	10	11.000	OneR	17	17.083	Logistic
2	4.083	CART	11	11.250	NaiveBayes	18	17.500	NaiveBayes
3	4.417	RIPPER	12	12.000	RBFNetwork	19	18.667	Kstar
4	5.167	LADTree	13	12.667	Logistic	20	19.667	VFI
5	6.000	LMT	14	13.245	NBTree	21	20.333	RandomTree
6	6.750	BayesNet	15	13.917	MLP	22	21.333	3NN
7	7.250	REPTree	16	14.083	Nnge	23	21.333	ZeroR
8	10.667	SMO				24	22.333	1NN
9	10.750	C4.5				25	24.667	HyperPipes

Table 3: Algorithms rankings according to the Friedman test for the prediction of voting intentions of a particular candidate; the best performer is BFTree.

Since the Friedman test rejects the null hypothesis, the Holm test [Holm, 79] was applied to determine which algorithms present significant differences on average sensitivity, with respect to the best perfo9rming algorithm (the BFTree). The results are shown in table 4.

i	Algorithm	Unadjusted p-value	p-value Holm test
1	HyperPipes	0.000	0.000
2	1NN	0.000	0.000
3	ZeroR	0.000	0.001
4	DecisionStump	0.000	0.001
5	LWL	0.000	0.005
6	3NN	0.000	0.005
7	RandomTree	0.000	0.005
8	VFI	0.000	0.006
9	Kstar	0.002	0.033

Table 4: Results of the Holm test when comparing the algorithms against the best performer (BFTree) on the prediction of voting intention for a particular candidate; only significant results are shown.

As can be seen, the Holm test rejects the null hypothesis for nine of the 25 algorithm under study. Regarding the first 16 algorithms in the ranking given by the Friedman test, the Holm test did not detect significant differences in their performance, and thus they are considered to be equally adequate for this task.

6.3 Results on the prediction of voting intentions for democrats and republicans

In the case of predicting a user voting intentions for some democrat or republican candidate, the results were higher than those related to predicting the voting intentions for a particular candidate. For this problem, performances above 75% were reached in all six experiments.

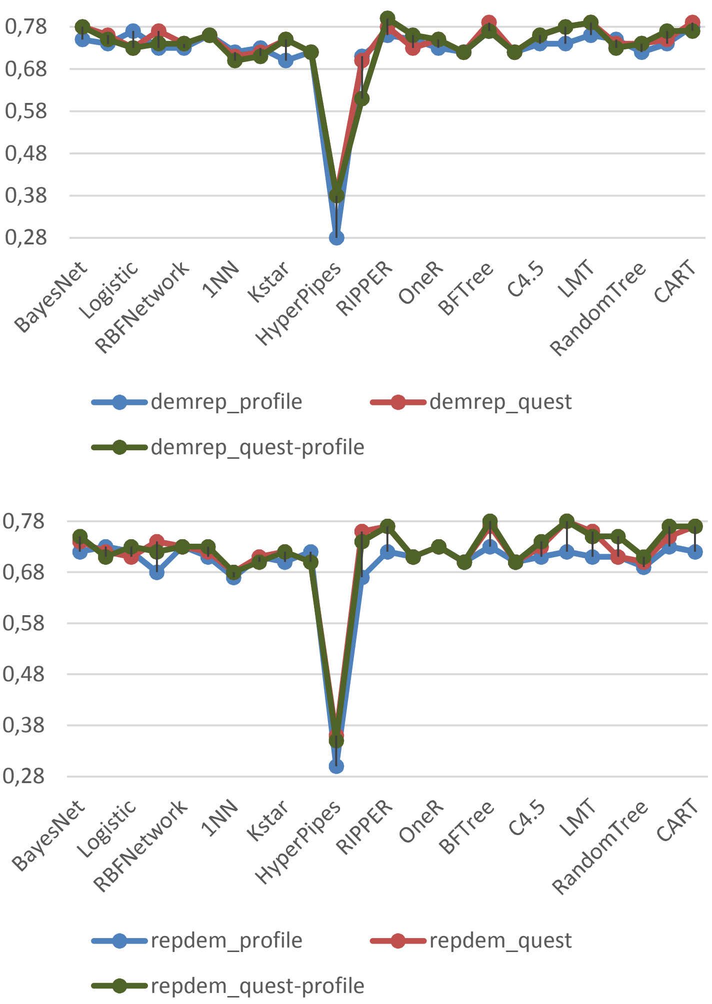

Figure 6: Average sensitivity per class of voting intentions for democrat or republican candidates.

Figure 6 shows the performance results obtained by the 25 classification algorithms, considering the average sensitivity per class (equation 1) for the task of predicting the voting intentions for democrat or republican candidates.

No.	Ranking	Algorithm	No.	Ranking	Algorithm	No.	Ranking	Algorithm
1	**3.3333**	**CART**	10	9.9166	SMO	17	17.0833	Logistic
2	3.9166	BFTree	11	11.75	RBFNetwork	18	17.5	NaiveBayes
3	4.6666	LMT	12	11.9999	DecisionStump	19	18.6666	Kstar
4	4.8333	LADTree	13	12.5	C4.5	20	19.6666	VFI
5	5.6666	REPTree	14	12.6666	LWL	21	20.3333	RandomTree
6	7.3333	NBTree	15	14.75	MLP	22	21.3333	3NN
7	7.4166	RIPPER	16	14.9166	Nnge	23	21.3333	ZeroR
8	8.0833	BayesNet				24	22.3333	1NN
9	8.3333	OneR				25	24.6666	HyperPipes

Table 5: Algorithms rankings according to the Friedman test for the prediction of voting intentions of a democrat or republican candidate; the best performer is CART.

i	Algorithm	Unadjusted p-value	p-value Holm test
1	HyperPipes	0.000	0.000
2	1NN	0.000	0.000
3	ZeroR	0.000	0.001
4	3NN	0.000	0.001
5	RandomTree	0.000	0.001
6	VFI	0.000	0.002
7	Kstar	0.000	0.006
8	NaiveBayes	0.001	0.015
9	Logistic	0.001	0.019

Table 6: Results of the Holm test when comparing the algorithms against the best performer (CART) on the prediction of voting intention for a democrat or republican candidate; only significant results are shown.

The prediction of voting intentions for a democrat or republican candidate according to variable *demcand* of the ANES online study has a maximum average sensitivity of 80%, while the prediction of voting intentions for a democrat or republican candidate based on the *repcand* variable of the questionnaire has a maximum average sensitivity of 78%. Such results are quite encouraging.

In order to decide which of the algorithms are more adequate to correctly predict the voting intentions in this scenario, the Friedman test was also applied [Friedman, 37; Friedman, 40] giving a value of $p = 6.01\mathrm{E}^{-11}$, again quite below the established significance level of $\alpha = 0.05$ for a 95% confidence. The rankings for the algorithms according to the Friedman test are shown in table 5, where the best classifier for this task is now the CART [Breiman et al., 84].

Now, the Holm test [Holm 79] was used to ascertain which algorithms present significant differences on their average sensitivities, with respect to the CART algorithm (the best performer). These results are shown in table 6.

As shown, the Holm test rejects the null hypothesis for nine of the compared algorithms. As for the first 16 algorithms appearing in the ranking given by the Friedman test, the Holm test detected no significant differences between their performances, and thus these methods are considered equally adequate for the task at hand.

6.4 Analysis of the attributes influence in predicting voting intention

Studying the impact that using some attributes or others in voting intention prediction is of particular interest. Figure 5 shows that using only the YouGov users profile data, the maximum average sensitivity reaches 55% for a democrat candidate and a 42% for a republican candidate, while using only attribute taken from the ANES study gives maximum average sensitivities of 70% and 55%, respectively. On the other hand, using all available attributes (i.e. combining the data from the YouGov profiles and the ANES questionnaire) offers respective maximum performances of 70% and 56%.

Prediction of a democrat candidate. Friedman p-value: $5.9\mathrm{E}^{-5}$		
Ranking	*Attributes*	*Holm p-value*
1.62	Questionnaire	-
1.66	Questionnaire + YouGov profile	0.8875
2.42	YouGov profile	0.0001
Prediction of a republican candidate. Friedman p-value: $6.96\mathrm{E}^{-4}$		
Ranking	*Attributes*	*Holm p-value*
1.64	Questionnaire	-
1.74	Questionnaire + YouGov profile	0.7237
2.62	YouGov profile	0.0011
Prediction of a democrat vote. Friedman p-value: 0.085		
Ranking	*Attributes*	*Holm p-value*
1.72	Questionnaire	-
1.94	Questionnaire + YouGov profile	-
2.34	YouGov profile	-
Prediction of a republican vote. Friedman p-value: 0.006		
Ranking	*Attributes*	*Holm p-value*
1.68	Questionnaire + YouGov profile	-
1.80	Questionnaire	0.6714
2.52	YouGov profile	0.0060

Table 7: Analysis of attribute impact on voting intention prediction.

Meanwhile, figure 6 indicates that using only features taken from the YouGov users profile gives maximum average sensitivity of 78% for democrat candidates and 73% for republican candidates (*demcand* and *repcand* variables in the questionnaire, respectively). On the other hand, using variables taken from the ANES study as input allows the classifiers to reach maximum performances of 79% and 78%, respectively. Finally, by combining both profile and questionnaire data the maximum performances are 80% for the *demcand* attribute, and 78% for the *repcand* variable. Yet, these results are insufficient to ensure the existence or absence of significant differences on the algorithms performances with respect to the feature sets used. For this, the Friedman test was used again, as well as the Holm test as the subsequent post-hoc test. Such results appear in table 7.

As can be seen, the Friedman test rejects the null hypothesis in every case, except for the prediction of voting intention for a democrat candidate versus not voting for a democrat candidate. In this instance (data set 7-9), it can be said with a 95% confidence that the YouGov user profile data is enough to predict whether a user intends to vote for a democrat candidate or not.

For the prediction of voting intentions for a republican candidate or not (data sets 10-12), the best results emerge when using both the YouGov user profile and the answers to the ANES study. However, there are no significant differences between using such combined data set and using only the questionnaire data.

In the analysis of voting intentions for a particular candidate, the YouGov profile alone is significantly worse for both scenarios, predicting the vote for a specific democrat candidate (data sets 1-3) and predicting the preference for one republican candidate (data sets 4-6). In both of these instances, the best results arise when the ANES questionnaire data is used as input; although no significant difference appear against using the combined data from the YouGov profile and the ANES study.

Such results support the assertion that the YouGov user profile includes attributes that differentiate the voting intention for a democrat candidate or not. However, these same attributes are not good enough by themselves to predict voting intentions for particular candidates (regardless of party), nor for predicting a voting preference for a republican candidate or not.

7 Conclusions and Future Work

In this paper, an extensive comparison between the performances of 25 classification algorithms was done, for the task of predicting voting intentions on the United States primary presidential elections for 2016. The data used for the experimental phase of this work was taken from the profiles of 1200 users of the YouGov OSN, as well as the answers they gave to an online study run by ANES, regarding said primary presidential elections. The 25 supervised classification algorithms were executed in the WEKA platform, using a stratified 5-fold cross validation scheme. Also, the experimental results obtained were validated in order to identify significant differences in performance by mean of a non-parametric statistical test (the Friedman test), and a post-hoc test (the Holm test).

The hypothesis testing analysis of the experimental results indicates that predicting voting intentions in favour of a democrat or republican candidate is simpler than predicting the particular candidate, given that the prediction performances for a democrat or republican candidate (best performances of 80% and 78%, respectively) are better than those given when predicting a specific candidate (70% for democrat candidates and 56% for republican candidates).

Also, in both of these two situations evaluated, nine of the 25 algorithms offered significantly poorer results, while 16 classifiers offered good results, which are also not significantly different between them. Curiously enough, 7 classifiers appear in both sets of nine poor performers, with HyperPipes being the worst in every experiment. For the two classes problem (democrat or republican candidate), the algorithm with best performance was the CART, while the best performance when predicting specific candidates (multiclass problem) was given by the BFTree classifier.

Another finding of particular interest is that the attributes taken from the YouGov user profile are enough to predict a democrat vote, while both the data taken from the YouGov profile and the ANES questionnaire are necessary to predict a voting preference for a republican candidate. On the other hand, the ANES study information gives the best result when predicting the voting preference for a particular candidate, whether the candidate is democrat or republican.

In near future the authors intend to apply other emerging methods to these same data sets [López-Yáñez et al., 14], as well as perform similar experimental comparison with other data sets, related to voting intention preferences.

Acknowledgements

The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, CIDETEC, and CIC), the CONACyT, and SNI for their economic support to develop this work. The authors also want to thank to the anonymous reviewers, whose valuable comments help to improve this paper.

References

[Aha and Kibler, 91] Aha, D., Kibler, D. “Instance-based learning algorithms”; Machine Learning, 6 (1991), 37-66.

[Alcalá-Fdez et al., 09] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, J.: “KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems”; Soft Computing, 13, 3 (2009), 307-318.

[Alcalá-Fdez et al., 11] Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: “KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework”; Journal of Multiple-Valued Logic and Soft Computing, 17, 2-3 (2011), 255-287.

[Atkeson et al., 97] Atkeson, C., Moore, A., Schaal, S.: “Locally weighted learning”; Artificial Intelligence Review, 11, 1-5 (1997), 11-73.

[Banai et al., 16] Banai, I. P., Banai, B., & Bovan, K.: “Vocal characteristics of presidential candidates can predict the outcome of actual elections”, Evolution and Human Behavior. In Press http://dx.doi.org/10.1016/j.evolhumbehav.2016.10.012

[Bartels, 00] Bartels, L. M.: "Partisanship and voting behavior, 1952-1996", American Journal of Political Science, 44(1) (2000), 35-50.

[Bicen and Uzunboylu, 13] Bicen, H., Uzunboylu, H.: The Use of Social Networking Sites in Education: A Case Study of Facebook; Journal of Universal Computer Science 19, 5 (2013), 658-671

[Bohrnstedt and Knoke, 82] Bohrnstedt, G.W. and Knoke, D. "Statistics for social data analysis", Peacock, Illinois (1982)

[Breiman et al., 84] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: "Classification and Regression Trees"; Wadsworth International Group, Belmont, California (1984)

[Brent, 95] Brent, M.: "Instance-Based learning: Nearest Neighbor With Generalization"; Hamilton, NZ (1995)

[Cameron et al., 16] Cameron, M. P., Barrett, P., & Stewardson, B: "Can social media predict election results? evidence from New Zealand", Journal of Political Marketing, 15 (2016), 416-432.

[Cleary and Trigg, 95] Cleary, J.G., Trigg, L.E.: "K*: An Instance-based Learner Using an Entropic Distance Measure"; In: 12th International Conference on Machine Learning (1995), 108-114.

[Cohen, 95] Cohen, W.W.: "Fast Effective Rule Induction"; In: Twelfth International Conference on Machine Learning (1995), 115-123.

[Demiroz and Guvenir, 97] Demiroz, G., Guvenir, A.: "Classification by voting feature intervals"; In: 9th European Conference on Machine Learning (1997), 85-92.

[Demšar, 06] Demšar, J.: "Statistical comparisons of classifiers over multiple data sets"; Journal of Machine Learning Research, 7 (2006), 1-30.

[Espinosa-Oviedo et al., 16] Espinosa-Oviedo, J. A., Vargas-Solar, G., Alexandrov, V., Castel, G.: "Comparing Electoral Campaigns by Analysing Online Data"; Procedia Computer Science, 80 (2016), 1865-1874.

[Fernández et al., 13] Fernández, A., López, V., Galar, M., Del Jesus, M. J., Herrera, F.: "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches"; Knowledge-based systems, 42 (2013), 97-110.

[Fernández-Navarro et al., 11] Fernández-Navarro, F., Hervás-Martínez, C., Gutiérrez, P. A.: "A dynamic over-sampling procedure based on sensitivity for multi-class problems"; Pattern Recognition, 44, 8 (2011), 1821-1833.

[Fisher et al., 2015] Fisher, S. D., Heath, A. F., Sanders, D., & Sobolewska, M.: "Candidate ethnicity and vote choice in Britain", British Journal of Political Science, 45(04) (2015), 883-905.

[Frank et al., 03] Frank, E., Hall, M., Pfahringer, B.: "Locally Weighted Naive Bayes"; In: 19th Conference in Uncertainty in Artificial Intelligence (2003), 249-256.

[Friedman, 37] Friedman, M.: "The use of ranks to avoid the assumption of normality implicit in the analysis of variance"; Journal of the American Statistical Association, 32 (1937), 674–701.

[Friedman, 40] Friedman, M.: "A comparison of alternative tests of significance for the problem of m rankings"; Annals of Mathematical Statistics, 11 (1940), 86–92.

[Friedman et al., 00] Friedman, J., Hastie, T., Tibshirani, R.: “Additive logistic regression: A statistical view of boosting”; Annals of Statistics, 28, 2 (2000), 337-407.

[García et al., 10] García, S., Fernández, A., Luengo, J., Herrera, F.: “Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power”; Information Sciences, 180, 10 (2010), 2044-2064.

[Garcia and Herrera, 08] Garcia, S., Herrera, F.: “An Extension on ‘Statistical Comparisons of Classifiers over Multiple Data Sets’ for all Pairwise Comparisons”; Journal of Machine Learning Research, 9 (2008), 2677-2694.

[Gayo-Avello, 12] Gayo-Avello, D. “No, you cannot predict elections with Twitter”. IEEE Internet Computing, 16(6), (2012), 91-94.

[Gayo-Avello et al., 11] Gayo Avello, D., Metaxas, P. T., & Mustafaraj, E. “Limits of electoral predictions using twitter”. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011), 490-493.

[Hall et al., 09] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: “The Weka Data Mining Software: An Update”; SIGKDD Explorations 11, 1 (2009), 10-18.

[Hastie and Tibshirani, 98] Hastie, T., Tibshirani, R.: “Classification by pairwise coupling”; The Annals of Statistics, 26, 2 (1998), 451-471.

[Holm, 79] Holm, S.: “A simple sequentially rejective multiple test procedure”; Scandinavian Journal of Statistics, 6 (1979), 65–70.

[Holmes et al., 01] Holmes, G., Pfahringer, B., Kirkby, R., Frank, E., Hall, M.: “Multiclass alternating decision trees”; In: ECML (2001), 161-172.

[Holte, 93] Holte, R.C.: “Very simple classification rules perform well on most commonly used datasets”; Machine Learning, 11 (1993), 63-91.

[John and Langley, 95] John, G.H., Langley, P.: “Estimating Continuous Distributions in Bayesian Classifiers”; In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo (1995), 338-345.

[Jung and Kazienko, 12] Jung, J. J., Kazienko, P.: “Advances on Social Network Applications”; J. UCS Special Issue. Journal of Universal Computer Science, 18, 4 (2012), 454-456.

[Keerthi et al., 01] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: “Improvements to Platt's SMO Algorithm for SVM Classifier Design”; Neural Computation, 13, 3 (2001), 637-649.

[Kohavi, 96] Kohavi, R.: “Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid”; In: Second International Conference on Knowledge Discovery and Data Mining (1996), 202-207.

[Landwehr et al., 05] Landwehr, N., Hall, M., Frank, E.: “Logistic Model Trees”; Machine Learning, 95, 1-2 (2005), 161-205.

[Le Cessie and van Houwelingen, 92] Le Cessie, S., van Houwelingen, J.C.: “Ridge Estimators in Logistic Regression”; Applied Statistics, 41, 1 (1992), 191-201.

[Little and Rubin, 89] Little, R. J., & Rubin, D. B.: “The analysis of social science data with missing values”. Sociological Methods & Research, 18(2-3) (1989), 292-326.

[López et al., 13] López, V., Fernández, A., García, S., Palade, V., Herrera, F.: "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics"; Information Sciences, 250 (2013), 113-141.

[López-Yáñez et al., 14] López-Yáñez, I., Sheremetov, L., Yáñez-Márquez, C.: "A Novel Associative Model for Time Series Data Mining"; Pattern Recognition Letters, 41 (2014), 23-33.

[Mattes and Milazzo, 14] Mattes, K., & Milazzo, C.: "Pretty faces, marginal races: Predicting election outcomes using trait assessments of British parliamentary candidates", Electoral Studies (2014), 34, 177-189.

[Metaxas et al., 11] Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. :"How (not) to predict elections". In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (2011), 165-171.

[Montag and Reuter, 15] Montag, C., Reuter, M.: "Internet Addiction. Neuroscientific Approaches and Therapeutical Interventions"; Springer International Publishing, Switzerland (2015)

[Niebert et al., 07] Niebert, N., Schieder, A., Zander, J., Hancock, R.: "Ambient Networks: Co-operative Mobile Networking for the Wireless World"; John Wiley & Sons, Ltd., Chichester, England (2007)

[Platt, 99] Platt, J. C.: "12 fast training of support vector machines using sequential minimal optimization"; Advances in Kernel Methods (1999), 185-208.

[Quinlan, 93] Quinlan, R.: "C4.5: Programs for Machine Learning"; Morgan Kaufmann Publishers, San Mateo, CA (1993)

[Sáez et al., 15] Sáez, J. A., Luengo, J., Stefanowski, J., Herrera, F.: "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering"; Information Sciences, 291 (2015), 184-203.

[Sang et al., 12] Sang, E. T. K., & Bos, J.: "Predicting the 2011 Dutch senate election results with twitter", In Proceedings of the Workshop on Semantic Analysis in Social Media (2012), 53-60.

[Seiffert et al., 14] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., Folleco, A.: "An empirical study of the classification performance of learners on imbalanced and noisy software quality data"; Information Sciences, 259 (2014), 571-595.

[Shi, 07] Shi, H.: "Best-first decision tree learning"; Hamilton, NZ (2007)

[Skoric et al., 12] Skoric, M., Poor, N., Achananuparp, P., Lim, E. P., & Jiang, J.: "Tweets and votes: A study of the 2011 Singapore general election", In 45th Hawaii International Conference on System Science (HICSS) (2012), 2583-2591.

[Smailovic and Podobnik, 16] Smailovic, V., Podobnik, V.: "Mining Social Networks for Calculation of SmartSocial Influence"; Journal of Universal Computer Science, 22, 3 (2016), 394-415.

[Sumner et al., 05] Sumner, M., Frank, E., Hall, M.: "Speeding up Logistic Model Tree Induction"; In: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (2005), 675-683.

[Sylvain, 02] Sylvain, R.: "Nearest Neighbor With Generalization". Christchurch, NZ (2002)

[Tahir et al., 12] Tahir, M. A., Kittler, J., Yan, F.: "Inverse random under sampling for class imbalance problem and its application to multi-label classification"; Pattern Recognition, 45, 10 (2012), 3738-3750.

[Tsakalidis et al., 15] Tsakalidis, A., Papadopoulos, S., Cristea, A. I., & Kompatsiaris, Y.: "Predicting elections for multiple countries using Twitter and polls". IEEE Intelligent Systems, 30(2) (20159, 10-17.

[Tumasjan et al., 10] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M.: "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". In Proc. of 4th ICWSM. AAAI Press, (2010), 178-185.

[Vluymans et al., 16] Vluymans, S., Tarragó, D. S., Saeys, Y., Cornelis, C., Herrera, F.: "Fuzzy rough classifiers for class imbalanced multi-instance data"; Pattern Recognition, 53 (2016), 36-45.

[Webster and Murphy, 08] Webster, L., Murphy, D.: "Enhancing Learning Through Technology: Challenges and Responses"; In Kwan, R., Fox, R., Chan, F.T., Tsang, P. (eds.). Enhancing Learning Through Technology: Research on Emerging Technologies and Pedagogies, World Scientific, Singapore (2008), 1-16.

[Witten and Frank, 05] Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques; Elsevier, Amsterdam (2005)

[You et al., 15] You, Q., Cao, L., Cong, Y., Zhang, X., & Luo, J.: "A Multifaceted Approach to Social Multimedia-Based Prediction of Elections", IEEE Transactions on Multimedia, 17(12) (2015), 2271-2280

[Zidan, 16] Zidan, K.: "New Social Network Platforms In 2016 That Is Not Facebook or Twitter"; Kindle Edition, Amazon Media, USA (2016)