Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
13/2 (2020), 89-96. DOI: http://dx:doi:org/10:21609/jiki:v13i2.860

MACHINE LEARNING FOR DATA CLASSIFICATION IN INDONESIA REGIONAL ELECTIONS BASED ON POLITICAL PARTIES SUPPORT

Muhammad Fachrie

Informatics Department, Faculty of Electrical and Information Technology, Universitas Teknologi Yogyakarta, Ringroad Utara, Sleman, Yogyakarta, Indonesia

E-mail: muhammad.fachrie@staff.uty.ac.id

Abstract

In this paper, we discuss the implementation of Machine Learning (ML) to predict the victory of candidates in Regional Elections in Indonesia based on data taken from General Election Commission (KPU). The data consist of composition of political parties that support each candidate. The purpose of this research is to develop a Machine Learning model based on verified data provided by official institution to predict the victory of each candidate in a Regional Election instead of using social media data as in previous studies. The prediction itself simply a classification task between two classes, i.e. 'win' and 'lose'. Several Machine Learning algorithms were applied to find the best model, i.e. k-Nearest Neighbors, Naïve Bayes Classifier, Decision Tree (C4.5), and Neural Networks (Multilayer Perceptron) where each of them was validated using 10-fold Cross Validation techniques. The selection of these algorithms aims to observe how the data works on different Machine Learning approaches. Besides, this research also aims to find the best combination of features that can lead to gain the highest performance. We found in this research that Neural Networks with Multilayer Perceptron is the best model with 74.20% of accuracy.

Keywords: *prediction, regional election, political party, machine learning, data mining*

Abstrak

Dalam artikel ini, kami membahas implementasi Machine Learning (ML) untuk memprediksi kemenangan kandidat pada Pemilihan Kepala Daerah di Indonesia berdasarkan data yang diambil dari Komisi Pemilihan Umum (KPU). Data tersebut terdiri dari komposisi partai politik yang mendukung masing-masing kandidat. Tujuan dari penelitian ini adalah untuk mengembangkan model Machine Learning yang berbasis pada data yang telah terverifikasi oleh lembaga resmi untuk memprediksi kemenangan masing-masing kandidat dalam pemilihan daerah alih-alih menggunakan data media sosial seperti dalam penelitian sebelumnya. Prediksi ini sendiri secara sederhana merupakan klasifikasi antara dua kelas data, yakni 'menang' dan 'kalah'. Beberapa algoritma Machine Learning diaplikasikan untuk menemukan model terbaik, yakni k-Nearest Neighbors, Naïve Bayes Classifier, Decision Tree (C4.5), dan Neural Networks (Multilayer Perceptron) yang masing-masing divalidasi menggunakan teknik 10-fold Cross Validation. Pemilihan algoritma-algoritma tersebut bertujuan untuk mengamati bagaimana data bekerja pada konsep matematika yang berbeda. Di samping itu, penelitian ini juga bertujuan untuk menemukan komposisi atribut-atribut terbaik yang mengarahkan pada pencapaian kinerja yang tertinggi. Kami menemukan dalam penelitian ini bahwa Neural Networks dengan arsitektur Multilayer Perceptron adalah model terbaik dengan akurasi sebesar 74,20%.

Kata Kunci: *prediksi, pilkada, partai politik, machine learning, data mining*

1. Introduction

Recently, Machine Learning (ML) is used in political field to do sentiment analysis from tweets, caption, or comments on social media, to predict the winner of general elections in several countries, such as United States [1-3], Germany [4], Italia [5] and some of Asian Countries such as Indonesia [6-8], Malaysia [9], India [9, 10], and Pakistan [9, 11]. All of these works are based on hypothesis that people's opinions about election candidates that are shared on social media represent the people choice in general election. This hypothesis is confirmed by [7] that election prediction that uses Twitter data gives good prediction accuracy, especially in developing country such as Indonesia. However, it has to be noticed that opinion shared on social media always have chance to be manipulated using any scenarios, e.g. fake account, paid buzzer, bot, etc.

Almost all previous research use sentiment analysis on Twitter data to perform the prediction. With this approach, the prediction can only be executed if there are enough number of tweets available on Twitter. Besides, it is required to collect more tweets in a certain period to increase the amount of dataset in order to reach a more precise prediction. Nevertheless, the twitter-based prediction gives good results on several cases, e.g. the 2014 presidential election in Indonesia was precisely predicted by [6] with only 0.61% of error rate compared to the real count published by General Election Commission (KPU). The similar performance also achieved by [7] with 0.62% of Mean Absolute Error (MAE) in the same election period. Another work by [9] that conducted the election prediction in Malaysia, India, and Pakistan, also prove that Twitter-based prediction is a good approach in predicting the election result with MAE lower than 5%.

Unfortunately, the problem comes over when there are not enough twitter data that can be collected to do the prediction, since the massive tweets about the election are usually shared when there are elections held in a big cities or provinces. Moreover, the collected tweets cannot be used immediately because it needs some text preprocessing, such as filtering the non-human (robot) tweets, filtering stop words, stemming, tokenizing, etc.

Hence, in this work we proposed an alternative approach by implementing Machine Learning to predict the general election result using multivariate dataset which contains a set of data served by General Election Commission (KPU). This approach can guarantee the validity of dataset that is used to train and test the Machine Learning model, so that the prediction is more reasonable compared to twitter-based model. Besides, it is not required to collect social media data in certain periods as usually used in tweet-based model which definitely requires more time to create a ready-to-use dataset.

2. Data and Methods

To analyze Machine Learning (ML) techniques on predicting the victory of candidates on regional elections, a set of data were collected from several credible sources. The dataset was used to train and evaluate the three different ML algorithms that is used in this research, i.e. Naïve Bayes Classifier, k-Nearest Neighbors, Decision Tree with C4.5 algorithm, and Neural Networks with Multilayer Perceptron architecture. The choice of these methods aims to get the comparison between basic statistical-based techniques, distance-based techniques, rule-based technique, and neural networks-based techniques.

2.1 Dataset

This research used primary dataset that was collected from several sources, mainly www.puskapol.ui.ac.id, pilkada2015.kpu.go.id, and pilkada2017.kpu.go.id. Moreover, other sources from www.wikipedia.org and some online newspapers were also gathered to complete the dataset.

The raw dataset contains of 1679 rows which represents 1679 candidates from 536 regional elections in Indonesia that was held from 2007 until 2018. It has 536 records of positive class which represents the elected candidate (symbolized by ‘1’) and 1143 records of negative class which represents the not elected candidate (symbolized by ‘0’). This is considered as imbalance dataset with ratio about 1:2. However, this gives impact to the classification performance.

There are total of 43 columns that contains of 1 label and 42 features. The label is the election result from each candidate in form of 1 and 0 which means ‘candidate wins’ and ‘candidate loses’ respectively. The 42 features are the support to the candidate and the percentage of seat acquisition in regional and central parliament from each political party. The support from each party that is presented in three different integers, i.e. 1, 0, and -1 which means ‘supports the candidate’, ‘neutral’, ‘do not support the candidate’ respectively. While the percentage of seat acquisition in regional and central parliament is calculated based on the number of parliament seats of each party divided by total number of seats on that parliament. Note that each regional parliament has different number of total seats to each other.

In this work, there are total of 12 political parties that exists in Indonesia during 2007 until 2018, i.e. ‘Partai Nasional Demokrat’ (Nasdem), ‘Partai Kebangkitan Bangsa’ (PKB), ‘Partai Keadilan Sejahtera’ (PKS), ‘Partai Demokrasi Indonesia Perjuangan’ (PDIP), ‘Partai Golongan Karya’ (Golkar), ‘Partai Gerakan Indonesia Raya’ (Gerindra), ‘Partai Demokrat’, ‘Partai Amanat Nasional’ (PAN), ‘Partai Persatuan Pembangunan’ (PPP), ‘Partai Hati Nurani Rakyat’ (Hanura), ‘Partai Keadilan dan Persatuan Indonesia’ (PKPI), and ‘Partai Bulan Bintang’ (PBB). In addition, there are three regional political parties in Aceh that took part on regional election in Aceh Province, i.e. ‘Partai Damai Aceh’ (PDA), ‘Partai Nasional Aceh’ (PNA), and ‘Partai Aceh’ (PA). Table 1 shows the description of dataset used in this research.

There are 12 elections in the dataset that only have single candidate. In result and discussion section, we will explain how these influence the classification performance.

TABLE 1
DESCRIPTION OF DATASET

Columns/ Features	Data Type	# of Columns	Role
Label	Integer {1, 0}	1 column	Label
Political Party Supports	Integer {1, 0, -1}	15 columns*	Feature
Percentage of Regional Parliament Seats	Real [0, 1]	15 columns*	Feature
Percentage of Central Parliament Seats	Real [0, 1]	12 columns	Feature

**It contains of 12 national parties and 3 regional parties*

2.2 Naïve Bayes Classifier

Naïve Bayes (NB) classifier implements the statistical approach in classifying data by calculating the probability of the data to be classified to certain class if given a set of training data. The probability is calculated using Bayes Theorem as shown in equation (1).

$$\boldsymbol{P(H|X)} = \frac{\boldsymbol{P(X|H)\,P(H)}}{\boldsymbol{P(X)}} \tag{1}$$

where $\boldsymbol{X}$ is evidence, and $\boldsymbol{H}$ is hypothesis. $\boldsymbol{P(H|X)}$ is probability that $\boldsymbol{H}$ is true if given evidence $\boldsymbol{X}$. $\boldsymbol{P(H|X)}$ is probability that evidence $\boldsymbol{X}$ is true if given hypothesis $\boldsymbol{H}$. While $\boldsymbol{P(X)}$ and $\boldsymbol{P(H)}$ is probability of evidence $\boldsymbol{X}$ and probability of hypothesis $\boldsymbol{H}$ respectively.

This work used dataset with continuous value for each attribute or feature. So, Gaussian probability was applied and Laplacian correction also implemented to avoid the zero probability.

2.3 k-Nearest Neighbors

k-Nearest Neighbors (k-NN) simply classifies a data based on its distance or similarity to other data around it. This method does not require learning process. Once the training data is collected, it can determine the class of new data based on the class of its neighbors. The number of neighbors determine the performance of k-NN algorithm.

In this research, distance measurement was calculated using Euclidean Distance which is given by equation (2). Besides, we used five different numbers of neighbors, i.e. 1, 3, 5, 7, and 9 to find the best model of k-NN.

$$\boldsymbol{d_{ij}} = \sqrt{\sum_{k=1}^{N}(x_{i_k} - x_{j_k})^2} \tag{2}$$

In equation (2), $\boldsymbol{d_{ij}}$ is the distance between data-i and data-j. $\boldsymbol{N}$ is the number of attribute of features, and x is a single tuple of data.

2.4 Decision Tree

Decision Tree classifier is a rule-based classification technique in Data Mining or Machine Learning. This rule-based model makes it easy to understand how a decision tree work to classify a data. Basically, a decision tree consists of a set of rules that are extracted from a training data using statistical approach. To determine which attribute should be placed as the root or as the branch of a node, it usually uses the information theory methods, such as information gain or gain ratio. The attribute that has the highest information gain or gain ratio value will be positioned as the root of the tree. Once a decision tree is successfully created, it can be viewed simply as IF-THEN rules.

This research applied C4.5 decision tree algorithm to classify the election result. This algorithm is selected due to its ability to deal with continuous data type that is used in this research. C4.5 utilizes Gain Ratio parameter instead of Information Gain that is usually used in ID3 algorithm, to select the most effective feature in classifying the data. The feature that has the highest value of Gain Ratio will placed as the root of decision tree. Gain Ration is calculated using equation (3) and (4).

$$\boldsymbol{GainRation(S,A)} = \frac{\boldsymbol{Gain(S,A)}}{\boldsymbol{SplitInfo(S,A)}} \tag{3}$$

where:

$$\boldsymbol{SplitInfo(S,A)} = \sum_{i=1}^{c} -\frac{|S_i|}{|S|}\boldsymbol{log_2}\frac{|S_i|}{|S|} \tag{4}$$

2.5 Neural Networks

Neural Networks has architectures that mimics the biological neural networks structure from the human, where there are a number of neurons that receive the input, then forward the input to other connected neurons, and finally when the signal is accepted by the neuron in the brain, it can conclude an output. In every connection between two neurons, there is an electrical stream which is the called as ‘weight’ in the context of ANN.

Multilayer Perceptron (MLP) was used in this research due to its simplicity and robustness in classifying data. An MLP architecture consists of three kinds of layers, i.e. an input layer, hidden layer (s), and one output layer, where each layer contains of several or a lot of neurons (node) depends on how complex the task given to it. MLP obtains the knowledge by learning the pattern of data using the Backpropagation algorithm, that is the very well-known learning algorithm for NN architecture that implements the concept of gradient descent. This algorithm works by optimizing the weight between connected neurons. Once it learns from the data, it will have a set of optimum networks weights. To obtain the best model of MLP, we design several scenarios to train and evaluate the system using including the number of hidden layer, the number of neuron in hidden layer, learning rate, maximum training iteration (epoch), and also momentum value.

2.6 Cross Validation

Cross Validation (X-Val) is a method that is used to evaluate and to validate the performance of our system in the context of classification. This method trains and tests the model using several different combinations of dataset. X-Val creates several combinations of separated training and testing data. The number of combinations is determined by the k-fold value, e.g. 3, 5, or 10. Figure 1 illustrate how X-Val works for k-fold = 3.

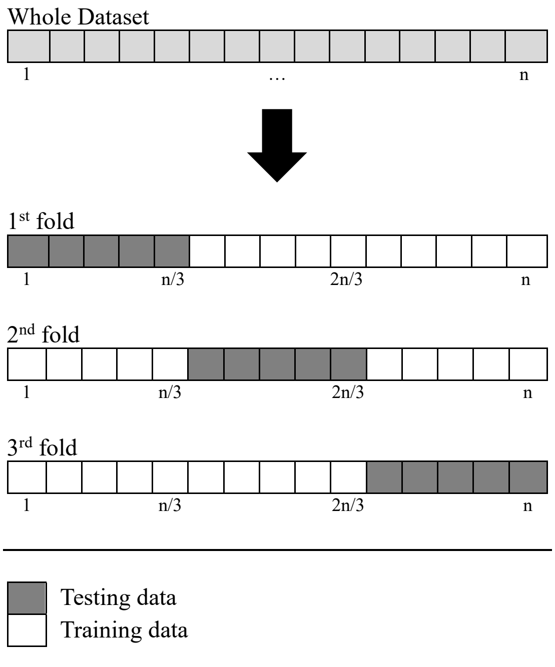

Figure 1. Illustration of how Cross Validation method creates several combinations of dataset using 3-folds.

X-Val is very beneficial for evaluating system with small size of dataset. It is recommended to use 10-fold of X-Val, so the system is trained and evaluated using many variations of data.

2.7 Primary Model Evaluation

In order to achieve the purpose of this research, three primary scenarios were applied to evaluate each ML algorithm, i.e. Naïve Bayes Classifier, k-Nearest Neighbors, Decision Tree (C4.5), and Neural Networks (Multilayer Perceptron). The first scenario evaluates the ML model using only 'Political Party Supports' attribute as mentioned in Table 1. This scenario has only 15 features.

The second scenario evaluates the model using 'Political Party Supports' and 'Percentage of Regional Parliament Seats' attributes. This scenario has the total of 30 features. The third scenario using all the features described in Table 1, hence it has the total of 42 features.

All these primary scenarios were applied to every ML algorithm using 10-fold of Cross Validation to find which model and which dataset gives the best performance.

2.8 Extended Model Evaluation

The dataset contains 12 elections with single candidate. This can be considered as outlier to our dataset. In Data Mining or Machine Learning area, outlier is capable to decrease the classification accuracy. Besides, there are three regional political parties which only exist in Aceh Province. This also can be considered as outlier. Hence, it needs to extend the primary model evaluation scenarios to find better model and better dataset.

The first extension excludes all the features from regional parties in Aceh Province, while the second extension deletes all tuples that has single candidate (no features are excluded). The third extension excludes all the features from regional parties in Aceh, also deletes all tuples that contains single candidate (combination of the first and second extension).

All the extended scenarios mentioned above was evaluated using the procedure used in the previous primary scenarios.

3. Results and Discussion

The experiment was performed using Rapidminer software, began with the three scenarios in the primary model evaluation to observe the performance of every algorithm against the original dataset, then continued to the extended model evaluation to optimize the model performance by modifying the original dataset with exclusion of some features (features that related to regional parties in Aceh) and some tuples (tuples with single candidate in election). Naïve Bayes (NB), k-Nearest Neighbors (k-NN), C4.5, and Multilayer Perceptron (MLP) were evaluated individually using the given dataset from each scenario. This task is performed using 10-fold Cross Validation technique.

3.1 Primary Model Evaluation Result

In primary model evaluation step, the first scenario which used 'Political Party Supports' attribute gained the highest best result among all scenarios with 70.06% of average accuracy from four ML algorithms as shown in Table 2. In this first scenario, there are only 15 features used in dataset compared to the second and the third scenario that uses 30 and 42 features respectively. The best result achieved in the 1st scenario is influenced by the accuracy of NB that is 5% - 6% better than accuracy achieved in 2nd and 3rd scenario as given in Table 2. NB performs pretty well in discrete or categorical data rather than continuous data. Whereas k-NN, C4.5, and MLP have pretty similar result in all scenarios since they was originally designed to deal with discrete and continuous data.

TABLE 2
RESULT OF PRIMARY MODEL EVALUATION

Methods	Model Accuracy		
	1st Scenario	2nd Scenario	3rd Scenario
NB	67.54%	62.54%	61.34%
k-NN	70.22%	71.11%	70.75%
C4.5	68.67%	67.48%	66.71%
MLP	**73.79%**	**73.97%**	**73.61%**
Average	**70.06%**	68.78%	68.10%

Based on Table 2, MLP becomes the best algorithm that achieved accuracies above 73% for all scenarios. This is not so surprising, given its ability to classify non linear data distribution. In this primary evaluation, MLP get the best performance in the second scenario with 73.97% of accuracy. Besides, k-NN that 'only' used simple distance-based measurement, achieved the second highest performance with larger than 70% of accuracies in all scenarios. However, the result in this experiment were obtained after several trial and error observations to find the right combination of parameters in order to get the best performance in every single algorithm. NB and C4.5 did not perform as good as k-NN and MLP in the primary evalutation.

3.2 Extended Model Evaluation Result

The primary model evaluation were extended to the next model evaluation to find out if regional parties features and single candidates record have influence to the model performance. The result shown in Table 3 answers the hypothesis that the regional parties gives impact to the model's performance. Every scenario in the first extended model evaluation omitted all the features that is related to regional parties in Aceh Province. This gives improvement to almost all algorithms in all scenarios. The significant improvement happens to NB in the second scenario with enhancement up to 8.93% of accuracy in all scenarios, while other algorithms just improve not more than 1%.

Regional parties features are dominated by zero values, because they only exist (have non zero values) in 102 out of 1679 records in dataset (only about 6% of the dataset). It means that the existence of these features create outliers that violate the data distribution. So that, the elimination of these features gives improvement to the models, especially to NB algorithm that uses probability-based measurement. The comparison of average accuracies between primary model evaluation and first extended evaluation is shown in Figure 2.

TABLE 3
RESULT OF THE FIRST EXTENDED MODEL EVALUATION
(WITHOUT REGIONAL PARTIES FEATURES)

Methods	Model Accuracy		
	1st Scenario	2nd Scenario	3rd Scenario
NB	72.07%	71.47%	69.86%
k-NN	70.93%	71.17%	70.40%
C4.5	68.73%	67.42%	66.83%
MLP	**73.97%**	**74.09%**	**73.97%**
Average	**71.43%**	71.04%	70.27%

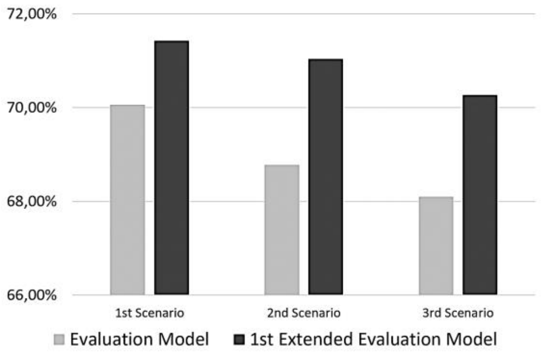

Figure 2. Comparison of average accuracies between primary evaluation model and 1st extended evaluation model which excluded regional parties in Aceh Province

The second step of extended evaluation model ignores the tuples which contain single candidate (regional parties features still exist in this evaluation). In contrary with the first extended evaluation, this step decreases the average accuracy of all scenarios except the third one as described in Table 4 and Figure 3. Unfortunately, the improvement on the third scenario is only 0.3% and the accuracy is still lower than 70%.

TABLE 4
RESULT OF THE SECOND EXTENDED MODEL EVALUATION (WITHOUT SINGLE CANDIDATE TUPLES)

Methods	Model Accuracy		
	1st Scenario	2nd Scenario	3rd Scenario
NB	67.13%	62.27%	61.73%
k-NN	69.28%	71.03%	70.85%
C4.5	69.19%	67.61%	67.01%
MLP	**73.90%**	**74.02%**	**74.02%**
Average	**69.87%**	68.73%	68.40%

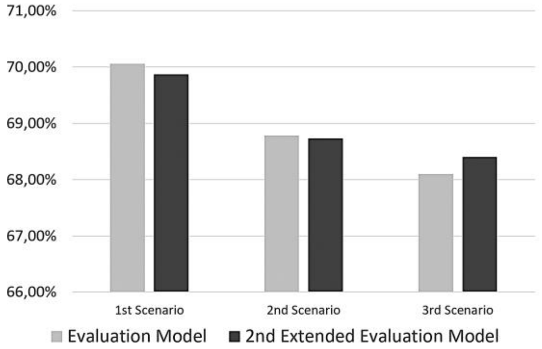

Figure 3. Comparison of accuracy between primary evaluation model and 2nd extended evaluation model which excluded tuples with single candidate.

However, MLP still gain the highest performance for all scenarios, even, little improvent occurs to the second and third scenario compared to the primary model evaluation. Hence, the elimination of single candidate records do not give good improvement to the system, even it results the worse model.

In the third extended model, we combined procedures in the first and the second one by excluding the regional parties features and eliminating all records that contain single candidate. The result which is described in Table 5 shows the improvement compared to primary model evaluation. But, this is still lower than the result in the first extended evaluation. The combination between excluding the regional parties features that increase the accuracy and eliminating the single candidate records that decrease the accuracy resulting the moderate accuracy to the model. It means that the existance of single candidate records in dataset contributes to create good classification model. Eliminating them do not result better classification model. The comparison of average accuracies from three scenarios between primary model evaluation and third extended evaluation is shown in Figure 4.

TABLE 5
RESULT OF THE THIRD EXTENDED MODEL EVALUATION (WITHOUT REGIONAL PARTIES AND SINGLE CANDIDATE)

Methods	Model Accuracy		
	1st Scenario	2nd Scenario	3rd Scenario
NB	72.22%	71.02%	70.73%
k-NN	69.46%	71.33%	70.97%
C4.5	68.86%	67.55%	66.89%
MLP	**74.20%**	**73.96%**	**74.02%**
Average	**71.19%**	70.97%	70.65%

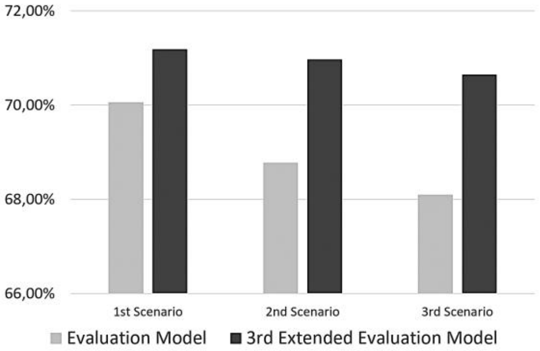

Figure 4. Comparison of accuracy between primary evaluation model and 3rd extended evaluation model which excluded regional parties features and tuples with single candidate.

In this third extended evaluation, MLP achieved highest individual accuracy with 74.20% in the first scenario. This accuracy is the highest among all model evaluation scenarios. In contrary, C4.5 may be the worst algorithm in this work. From all extended evaluation scenarios, the highest accuracy that it can reached is only 69.19% in the first scenario of the second model evaluation as shown in Table 4. NB also get improvement compared to previous extended evaluation with 72.22% of accuracy which is the highest among its accuracies from all evaluation scenarios. The comprehensive comparison among all algorithms in all model evaluation scenarios is delivered in Table 6.

Based on Table 6, MLP with the 1st scenario on the 3rd Extented model evaluation achieved the best performance with 74.20% of accuracy. It means that dataset that uses ‘Political Party Supports’ in the 1st scenario by excluding the regional party features and single candidate tuples, is the best model. This result was obtained due to the distribution of the data in the 1st scenario is well separated, so that ML algorithms used in this work, especially MLP, can easily classify the data into two different classes.

TABLE 6
COMPARISON AMONG ALL ALGORITHMS IN EVERY MODEL EVALUATION SCENARIOS

Model Eval.	Algorithms			
	NB	k-NN	C4.5	MLP
Prim. Model Eval.	67.54%	70.22%	68.67%	73.79%
	62.54%	71.11%	67.48%	73.97%
	61.34%	70.75%	66.71%	73.61%
1st Ext. Model Eval.	72.07%	70.93%	68.73%	73.97%
	71.47%	71.17%	67.42%	74.09%
	69.86%	70.40%	66.83%	73.97%
2nd Ext. Model Eval.	67.13%	69.28%	**69.19%**	73.90%
	62.27%	71.03%	67.61%	74.02%
	61.73%	70.85%	67.01%	74.02%
3rd Ext Model Eval.	**72.22%**	69.46%	68.86%	**74.20%**
	71.02%	**71.33%**	67.55%	73.96%
	70.73%	70.97%	66.89%	74.02%
Avg.	67.49%	70.63%	67.75%	**73.96%**

TABLE 7
STANDARD OF DEVIATION FROM EACH FEATURES ON DATASET

Political Party	Support	Central	Regional
Nasdem	0.833	0.013	0.046
PKB	0.814	0.007	0.054
PKS	0.808	0.006	0.044
PDIP	0.870	0.005	0.083
Golkar	0.827	0.006	0.062
Gerindra	0.884	0.018	0.040
Demokrat	0.889	0.030	0.054
PAN	0.862	0.001	0.056
PPP	0.704	0.002	0.046
Hanura	0.827	0.002	0.043
PKPI	0.532	0	0.0038
PBB	0.625	0.001	0.033

The additional features added to the dataset, i.e. ‘Percentage of Regional Parliament Seats’ and ‘Percentage of Cental Parliament Seats’, do not give positive impact to the model accuracy. This happens due to the distribution of data becomes more complex due to the large number of features. Besides, the value of ‘Percentage of Regional Parliament Seats’ and ‘Percentage of Central Parliament Seats’ features have small standard of deviation value as given in Table 7. It means that the value of data in both of additional features are likely uniform, so it can not give significant improvement to the model accuracy.

However, the highest accuracy obtained in this work is not followed by good recall. For example, as shown in confusion matrix in Table 8, MLP with the 74.20% accuracy in the third extended model evaluation only get class recall 38.29% for the data with label ‘1’. It happens due to imbalance class in dataset that has ratio about 1:2 between class ‘1’ and ‘0’. It can be explored and improved in the next research. In this work, we also find the other factor that also influence the model accuracy, that is the contradiction of some data that has similar features value but different class or label. Based on our observation, there are 26 out of 1679 data that has contradiction.

TABLE 8
CONFUSION MATRIX OF MLP FROM THE FIRST SCENARIO OF THE THIRD EXTENDED MODEL EVALUATION

	True 0	True 1	Class Precision
Predicted 0	1036	324	76.18%
Predicted 1	106	201	65.47%
Class Recall	90.72%	38.29%	

4. Conclusion

All scenarios have been completed with various results. But, there is still no classification model that achieve very good performance above 80% of accuracy. Yet, this is still preliminary research that can be explored in the next researches. At least, good performance with more than 74% of accuracy has been achieved by only using the features of political parties’s support to candidates.

To sum up, this research has delivered the experiment that implements Machine Learning (ML) algorithms to predict the victory of candidates in regional elections. We have compared some ML algorithms that has different approach in classification the data. However, in this work, the best model that successfully obtained is in the first scenario of the third extended evaluation model. This model uses MLP as the classifier and dataset that excludes regional parties features and eliminates single candidate records. The highest accuracy is 74.20% which is achieved by MLP, while the lowest accuracy is 61.34% which is obtained by NB. Besides, rule-based and distance-based algorithms, i.e. C4.5 and k-NN is not suggested to be used for this case due to the lack of performances. Moreover, the dataset used in this work has imbalance class with ratio 1:2, hence it needs additional preprocessing step to fix the imbalance problem in the next research.

At last, this approach is suitable to be implemented on regional election in other country that uses multiple party in their political system. But, different performance may be occur due to different society, demography and political situation in every country.

References

[1] M. Zolghadr, S.A.A. Niakin, & S.T.A. Niaki, “Modeling and forecasting US presidential election using learning algorithms,” *Journal of Industrial Engineering International*, vol. 14, pp. 491-500. 2017.

[2] K. Jahanbakhsh & Y. Moon, “The Predictive Power of Social Media: On the Predictability of U.S. Presidential Elections using Twitter,” 2014, [Online]. Available: https://arxiv.org/pdf/1407.0622.pdf

[3] E. Tunggawan & Y. E. Soelistio, “And the Winner is ... : Bayesian Twitter-based Prediction on 2016 U.S. Presidential Election,” in *Proceeding of IC3INA 2016*, 2016, pp. 33-37.

[4] E.F.M. Araujo & D.Ebbelaar, “Detecting Dutch Political Tweets: A Classifier based on Voting System using Supervised Learning, ” in *Proceeding ICAART 2018*, 2018, pp. 462-469.

[5] M. Coletto, C. Lucchese, S. Orlando, & R. Perego, “Electoral Predictions with Twitter: a Machine-Learning approach,” in *Proceeding IIR ,* 2015.

[6] M. Ibrahim, O. Abdillah, A.F. Wicaksono, & M. Adriani, “Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in A Twitter Nation,” in *Proceeding of ICDMW 2015*, 2015, pp. 1348-1353.

[7] N.D. Prasetyo & C. Hauff, “Twitter-based Election Prediction in the Developing World,” in *Proceeding of ACM Conference on Hypertext & Social Media 2015,* 2015, pp. 149-158.

[8] W. Budiharto & M. Meiliana, “Prediction and Analysis of Indonesia Presidential Election from Twitter Using Sentiment Analysis,” *Journal of Big Data*, vol. 5, pp. 1-10. 2018.

[9] K. Jaidka, S. Ahmed, M. Skoric, & M. Hilbert, “Predicting Elections from Social Media: A Three-country, Three-method Comparative Study,” *Asian Journal of Communication*, vol. 29:3, pp. 252-273. 2018.

[10] B. Bansal & S. Srivastava, “On Predicting Elections with Hybrid Topic Based Sentiment Analysis of Tweets, ” in *Proceeding of ICCSCI 2018*, 2018, pp. 346-353.

[11] M.A. Razzaq, A.M. Qamar, & H.S.M. Bilal, "Prediction and analysis of Pakistan election 2013 based on sentiment analysis," in *Proceeding of ASONAM 2014*, 2014, pp. 700-703.