Data Analysis and Feature Selection on Tax Comments Using Machine Learning Algorithms

[1]Shaik Naazreen Sulthana, [2]Dr. K. Venkata Rao

[1]M. Tech Scholar, [2]Professor,
Department of Computer Science and Systems Engineering,
Andhra University College of Engineering (A), Visakhapatnam, AP, India.

Abstract: The tax plays a crucial role for the contributions of the country's economy and its development. This research uses the data source as big data analysis due to the rapid growth of data in social media. In processing tax comments Facebook and Twitter were used as a source from which the data is derived. The results of opinions in the form of public sentiment in part of service, website system and news can be considered to improve the quality of tax services. In this paper, text mining is done through the various phases of text processing, feature selection and classification with Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN). The main purpose of this paper is to provide a comparison of some commonly used classification algorithms under same conditions. This type of comparison helps to provide the algorithms with more accurate result. The application of Artificial Neural Network (ANN) on Tax Comments which results prediction, is the main focus of this project. In doing so, we identify the learning methodologies utilized, data sources and specific challenges of predicting Tax Comments results. The predicted values for the classifiers were evaluated and the results were compared.

Keywords: Machine-learning Techniques, Minmax Scalar, Data Mining, SVM, KNN, ANN.

1.Introduction

The economy of the country is supported strongly by the number of the country's State Budget. Sources of State budget include Taxes, Non-Tax State Revenues and grant receipts both from within and outside the country. Taxes became the largest source of contributions to the State treasury, accounting for 85.6% of all State budget [2]. The types of taxes that are often charged are the income tax, Value-Added Tax, Sales tax in luxury goods and other taxes. In order to realize the target of the States budget in the coming years, it is necessary to have various efforts and cooperation to improve the tax system both from the government and society side. Efforts made by the Directorate General of Taxation are the online complaint services and the development of tax applications including the State Acceptance Modules such as e-billing, e-invoice and e-filing facilities. The growth of internet and social networking today has made it easy for the people to express their opinions. Complaints submitted by the public through Facebook and Twitter can be extracted into considered in the evaluation of the quality of tax services. The use of social media by the users has encouraged the increase of unlimited textual information, so that there is a need to utilize textual data to be presented without reducing the value of the information.

In doing text processing it required the use of classification methods such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Networks (ANN). This paper proposed the comparison of results some commonly employed Classification algorithms under same condition through SVM, KNN and ANN algorithms and to prove that Artificial Neural Networks (ANN) classification algorithm with feature selection has better rate of accuracy, recall and precision when compared with other classification algorithms. As the Feature selection is used to select the relevant feature of dataset in order to get the better performance of ANN as a classifier. The metrics used to calculate the result are Accuracy, Recall and Precision. In this study the results are based on time period and type of tax data namely services, websites and tax comments. For further research, information generated from this can be used as support services for future policies. Machine learning provides methods techniques and tools which help to learn automatically and to make accurate predictions based on the past observations.

This paper is organized as follows. In section II, we review the related research study in the fields of data Analysis, feature selection. In section III, presented a description of various processing's carried out on the dataset. The performance evaluation and its measures are introduced in section IV. Section V highlights the evaluated methodology with a discussion of experiment results and comparison with previous studies using some commonly employed algorithms on same data under same conditions. Finally, in section V, the conclusion discussed.

2. Related Study

In previous studies, Support Vector Machine (SVM) approach is widely used as machine learning algorithm for data analysis on tax comments. The researchers had proposed various researches related to text processing using certain methods to optimize performance levels including improving accuracy and reducing the error rate in classification.

Existing researches are largely based on supervised learning approach such as Support Vector Machine (SVM). Mihuandayani,et al.[1] proposed that Text mining in this research is obtained through the stages of text processing, feature selection using Information Gain and classification using SVM. Sheshasaayee, A et al. [4] proposed SVM based classifier to obtain better degree of accuracy in classification. Wang, Set al. [5] proposed that SVM is a method that overcomes over machine learning, but one of the problems with text classification is number of attributes used on a dataset. Ahmad et al. [10] proposed that the SVM performance that results in values that depend on the dataset as well as the ratio of training data and testing data. Xu, T., Peng, Q et al. [6] proposed that the existing attributes must be selected with right algorithm to get the better accuracy. Sulistiani, H et al. [7] proposed that feature selection is based on the subsets that works by

minimizing features that are not relevant to classification. Croft et al. [8] proposed that One of the most widely used feature selection criteria for classification applications is the Information Gain. Lu, H., Setiono et al. [3] proposed that ANN applied to classification based on the extracted rule has a low error rate usually.

3. Methodology

Classification Algorithms

Supervised learning algorithms were adopted for analyzing real time dataset and to predict the performance. Accuracy improvement is the main motivation for different classification algorithms. Classification is technique to categorize our data into a desired and distinct number of classes where we can assign label to each class. An attempt to draw some conclusion from observed values is done by classification model. If one or more inputs are given to a classification model then it will try to predict the value of one or more outcomes. Various classification models are used in the area of Machine learning and knowledge discovery. Different classification methods were Naive Bayes Classifier, K-Nearest Neighbor [KNN], Support Vector Machines [SVM] etc., Each method has its own variety of algorithms. Various algorithms were used to predict the accuracy of the dataset. The classification methods used here are SVM, KNN and ANN.

3.1 K Nearest Neighbors (KNN)

One of the most basic yet essential classification algorithm in machine learning is K-Nearest Neighbors which belongs to the supervised learning. Based on the points that are most similar to it the KNN model classifies data points and the output depends on whether *k*-NN is used for classification or regression:

- In k-NN classification, output is a class member. An objects classification is done by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is assigned simply to the class of that single nearest neighbor.

A unusual feature of the *k*-NN algorithm is that it is sensitive to the local structure of the data.

KNN Algorithm

Step 1: Load the training and test data,
Step 2: Choose the value of k.
Step 3: For each point in test data:

- find the Euclidean distance to all the training data points.
- store the Euclidean distance in a list and sort it.
- Choose the first k points.
- Assign a class to test point based on the majority of classes present in chosen points

Step 4: End.

3.2 Artificial Neural Network (ANN)

The ANNs is based on the belief that working of human brain by making the right connections, can be imitated using silicon and wires as living neurons and dendrites. As the neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data it has taken. The result of these operations is passed to other neurons. The output at each node is called its node value**.** The following Fig 2 shows a simple ANN –

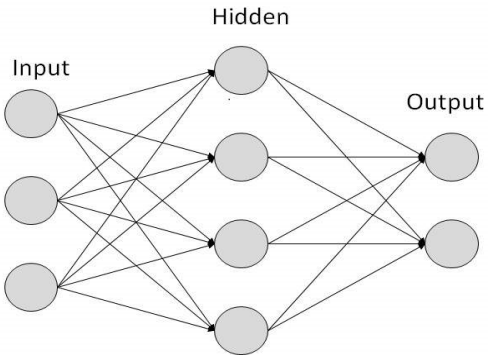

Fig 1. Artificial Neural networks

The neural network consists of the following 3 layers:

- **Input Layer**: The main purpose of the input layer is to receive the attributes for each observation. The number of nodes in the input layer is equal to the number of variables. Input layer represents the pattern of network which communicates with one or more hidden layers
- **Hidden Layer**: The main purpose of hidden layer is it performs computations and transfer information from input layer to output layer.
- **Output Layer**: The hidden layer links to the output layer and the output layer receives connections from hidden layers or from input layers. The active nodes at the output layer combines and change the data to generate the output.

4. Evaluation and Results

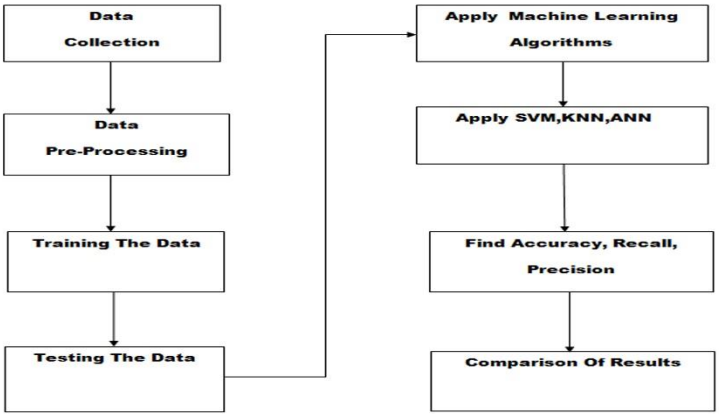

Fig.2 shows our proposed methodology

4.1 Data Collection

The dataset used in processing tax comments is derived from Facebook and Twitter as a source of data. The results of opinions in the form of public sentiment in part of service, website system, and news can be used as consideration to improve the quality of tax services.

4.2 Data preprocessing

Once data is gathered, it needs to be preprocessed, cleaned, constructed, and formatted in a style that ANN comprehends and is able to work with. ML tools should be used to analyze collected real-time data. Machine learning algorithms works better when features are on relatively similar scale.

4.2.1 Minmax Scalar

Minmax scalar is used to preprocess data in machine learning. For every value in feature, Minmax scalar subtracts the minimum value in the feature and then divide by the range. This range is the difference between the original maximum and original minimum. Minmax scalar preserves the shape of original distribution. Minmax scalar does not reduce the importance of the outliers.0 to 1 is the default range for the feature returned by minmax scalar. A minmax scaling is typically by the equation given below

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad \text{Eq.}$$

where, x is the original value

x' is the normalized value

4.3 Training and Testing Data

Training and testing are the two important concepts of machine learning.

Training Dataset :- A subset to train a model.

Testing Dataset :- A subset to test the trained model.

4.4. Results

The proposed model needs to be trained and tested by altering ANN parameters so that correctness can be obtained. In addition, we consider that the model's accuracy is maximum. From the collected data, 70:30 will be used to train and test the model, respectively. The data set was segregated into two parts, one part is used as training data set to produce the prediction model, and the other part is used as test data set to test the accuracy of our model. Two learning performance evaluators are included in PYTHON. The training data set consists of feature values as well as classification of each record. These test data were not used for training purpose. Testing will be carried out until every data appeared in the test set. Since a clear decision could not be made during the first set of data set, the train phase was conducted again for a different data to analyze the performance of the various algorithms. Computation of a confusion matrix done for every test [9]. Usually it is very difficult to predict large dataset due to data randomness. Hence testing for larger datasets would give us the flexibility to analyze each algorithm's real effectiveness in prediction. The results of the various classification algorithms are given below.

Fig 3: Precision values for the algorithms

Precision: Precision refers to the percentage of your results which are relevant that are correctly classified by the algorithm
.

Fig 4: Recall values for the algorithms

Recall: Recall is the position of correctly predicted instruction versus total number actual instructions.

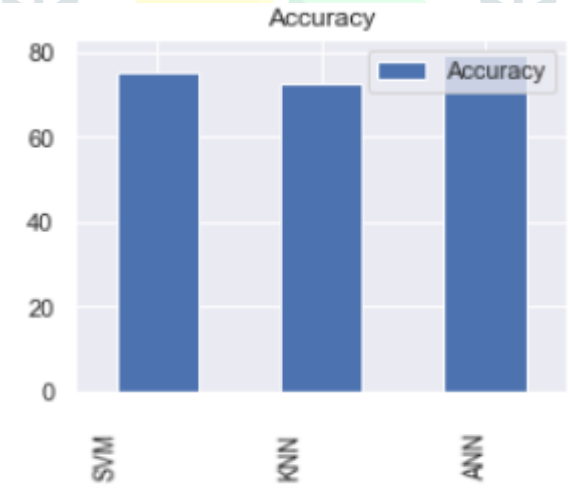

Fig 5: Accuracy values for the algorithms

Accuracy: Accuracy is the number of correct predictions made divided by total number of predictions.

4.4. Comparison of Models

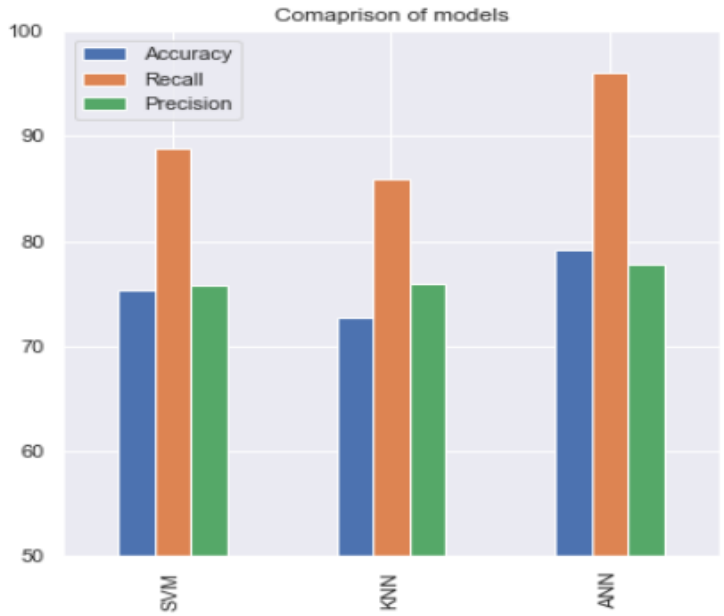

Fig 6: Comparison of SVM, KNN, ANN algorithms.

The result showing accuracy of various Classification Algorithms**:**

Table.1 Accuracy of Models

SNO	CLASSIFICATION ALGORITHMS	ACCURACY
1	SUPPORT VECTOR MACHINE [SVM] ALGORITHM	75
2	K NEAREST NEIGHBOURS [KNN] ALGORITHM	72
3	ARTIFICIAL NEURAL NETWORK [ANN] ALGORIHM	78

5. Conclusion

This paper presents results in the field of data classification obtained with three different classification methods of Python tool like SVM, KNN, ANN. ANN algorithm was selected in each method to predict the accuracy of dataset. These techniques have been implemented using PYTHON and the independent trained models were generated. The performance of the learned models was evaluated based on their predictive accuracy and ease of learning. Based on the experimental results the ANN classification accuracy has found to be better using function classifier than other two classifications. From the above results it has been observed that the ANN algorithm plays a major role in determining better classification accuracy in the dataset. Thus, from all perspectives, the ANN could be considered as the most efficient.

REFERENCES

1. Mihuandayani, Utami, E., & Luthfi, E. T. (2018). Text mining based on tax comments as big data analysis using SVM and feature selection. 2018 International Conference on Information and Communications Technology (ICOIACT). doi:10.1109/icoiact.2018.8350743

2. Admin. (2017, October 15), State Revenue Realization - Ministry of Finance of the Republic of Indonesia. Available online:www.bps.go.id

.

3. Lu, H., Setiono, R. Liu, H. NeuroRule: A Connectionist Approach to Data Mining. 2017. Available online: arXiv:1701.01358v1 [cs. LG].

4. Sheshasaayee, A. & Thailambal G. Comparison of Classification Algorithms in Text Mining. International Journal of Pure and Applied Mathematics 2017, Vol. 116 No.22 pp 425-433.

5. Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Systems with Applications, 38(7), 8696–8702. doi:10.1016/j.eswa.2011.01.077.

6. Xu, T., Peng, Q., & Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. Knowledge-Based Systems, 35, 279–289. doi:10.1016/j.knosys.2012.04.011.

7. Sulistiani, H., & Tjahyanto, A. Comparative Analysis of Feature Selection Method to Predict Customer Loyalty. Journal of Engineering, Vol. 3, No. 1, 2017 (eISSN:2337-8557).

8. Croft, W.B, Metzler D, and Strohman T. 2015. Search Engines: Information Retrieval in Practice. Pearson Education.

9. Dedhia, C and Ramteke, J. Ensemble model for Twitter Sentiment Analysis. International Conference on Inventive Systems and Control 2017.

10. Ahmad, Munir, & Shabib Aftab. Analyzing the Performance of SVM for Polarity Detection with Different Datasets International Journal Modern Education and Computer Science. DOI: 10.5815/ijmecs.2017.10.04.