Empirical Analysis of Selected Supervised Machine Learning Algorithms on Twitter Data

Dipak Patil[1] , Poonam Katare[2], Parag Bhalchandra[3], Aniket Muley[4]
[1]Department of Comp. Sci., L.K.R.B.Patil Mahila Mahavidyalaya, Islampuer, Sangali, MS, India
[2]MCA Department, PCCOE, Pune, India
[3]School of Computational Sciences S.R.T.M.University, Nanded, MS, India,
[4]School of Mathematical Sciences S.R.T.M.University, Nanded, MS, India.

Abstract: This paper demonstrate an investigational research work carried out Twitter social network site in order to do comparison of selected supervised learning algorithms with existing standard methods. The experimental work is done on python platform. The comparison of algorithms is made with respect to precision, accuracy, recall and F1 scores.

Keywords: Data Mining, Social Network Analysis, Classification, Accuracy

I. Introduction

The usage of www and internet has given tremendous scope for interactions between people mainly in personal and social contexts. The social networks have been established for persons to share thoughts, dreams, photos and videos, posts, and to notify others about happenings in their spheres with other with people in their contacts or groups. Such networks have been moderately studied from the point of view of understanding interactions between people and their patterns as well as impacts [1]. Since social animals like humans are involved in the interactions, social networks have diverse groups according to the taste and other qualities. The social network information of connected people is known to everybody but the relationship within the groups is hidden. Discovery of this is complex as relationship is invisible. That is why people usually deploy analytics, mining methods, and machine learning methods to find out patters, behaviors, sentiments, group forming tendencies, likes, dislikes, etc. This is called as social network analysis which primarily works for discovery of invisible patterns. The link analysis is popular method of social network analysis which uses network graphs to find out patterns in the relationship of the objects being considered [6]. Such network graphs help us to classify, cluster, predict, associate, etc all aspects of the analytics to be carried out. Desired analytics are implemented on such graphs by defining metrics and models on the network graphs [2,4]. To demonstrate of experimentations, in this study we have used datasets from Twitter [3]. It consists of 973 networks with 81306 nodes and 1768149 edges. The standard definition of all attributes and variables of this dataset are as mentioned in [3].

II. Experimental Set up

Standard data cleansing procedure [3] is adopted to remove duplicates, missing values, redundancy etc from the dataset. Later attempts are made to devise out the desired classification models. Appropriate training is given and the model is fitted over the dataset. Since ultimate aim is to compare the performance issues, we have found out parameters like f1-Score, accuracy, recall, etc [3]. Experimentations are done and comparisons are also made as discussed in below sections. The overall research methodology is as shown in below figure 1,

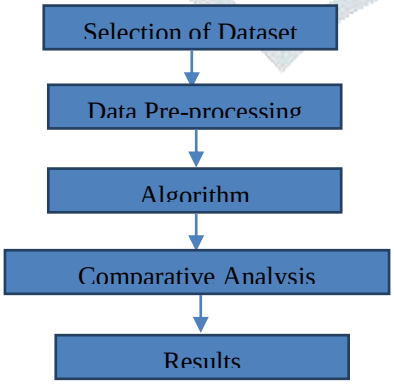

Fig. 1: Research Methodology

Various Classification Algorithms are present in Supervised Machine Learning. We have selected K-Nearest Neighbors Algorithm (KNN), Logistic Regression Classifier Algorithm, Random Forest Classifier Algorithm and AdaBoost Classifier Algorithm for comparison [1].

1. **K-Nearest Neighbors Algorithm (KNN):** The KNN algorithm works on nearness or proximity [1,5]. It works on distance based theory and usually Euclidean distance is used for core calculations. It considers similarity of training features as the main processing style so as to predict new data points. We have adopted its working as enlisted in [7].

2. **Logistic Regression Classifier Algorithm**: It is based on the use linear equation to predict data point values between a space having boundaries from infinite negative to infinite positive. The sigmoid function is used for this prediction [1] as shown below,

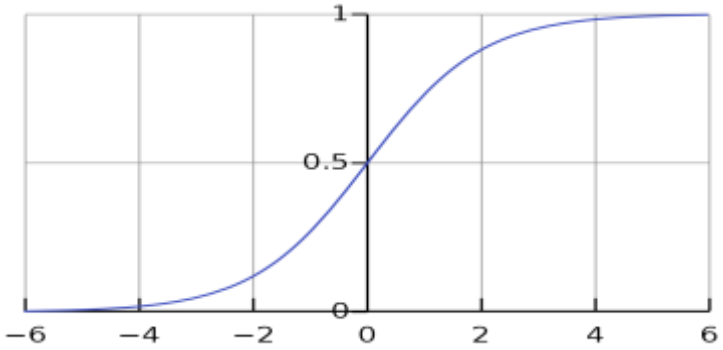

Fig.2: Standard graph for Sigmoid Function with x as input and y as output

If input is x then, from the graph, the output asymptote positive is at y=1 and asymptote negative is at y=0. While predicting the output classification class, as the likelihood gets closer to 1, there is gain in the confidence of the model that the observation is in class 1.

3. Decision Tree Classifier Algorithm: A Decision tree has nodes as classification choices and the edges depict the conditions, which if become true, we traverse down the edge to look for class output or further classifications [1,5,7]. It's more or less like a flowchart with number of logical decisions to be taken from root node that is starting point to end node that is leaf. The leaf nodes are last nodes to be traversed and the decision tree ends at leaf nodes which particularly represent classification. Trees algorithm is graspable, and mimic human decision making if significant information is in hands[7].

4. Random Forest Classifier Algorithm: A random forest approach creates lot of decision trees during entire classification process, specifically, every sample has a decision tree [1,5,7]. Later on prediction work begins from each of the trees and voting is done to find most suitable one. Maximum voted results are treated as final classification. We stick to following working methodology of random forest [7],

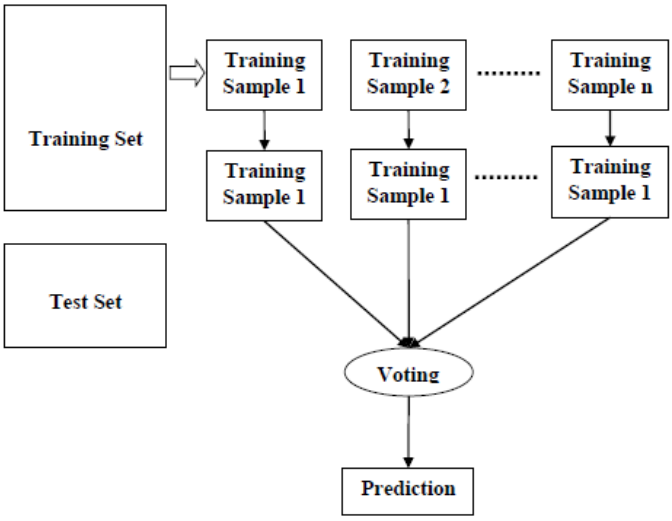

Fig.3: Random Forest work style

5. AdaBoost Classifier Algorithm: AdaBoost is one of the most successful boosting ensemble methods where weights are given to the dataset instances [1,5]. During model construction, base_estimator. Here, base_estimator module is prepared first. Then onwards, training model is constructed using incremental approach and using weak learners sequentially. We have followed the methodology of AdaBoost as listed in [7].

III. Results and Discussions using algorithm evaluation:

All machine learning Classifier Algorithms are implemented on Twitter dataset. During comparison work, we stick up to following rules,

1. Evaluation of the model for classifier is compulsorily done to check whether output is satisfactory or not?
2. Performance accuracy of the model is understood as "how correctly and incorrectly instances are classified out of total number of instances". The result is expressed as a percentage and is shown in the table 1.
3. Use of Confusion matrix for describing performance of the model. It has some attributes like False Positive (FP), False Negative (FN), True Positives (TP) and True Negatives (TN).

4. Total sample count is sum of all FP, FN, TP and TN.
5. Use of Accuracy parameter to record correctness of the model, which is TP+TN / Total sample count.
6. Precision is taken as TP+FP/ Total sample count.
7. Recall is taken as TP+FN/ Total sample count.
8. F1 Score is harmonic mean of Precision and Recall as they both deal with relevance. It is best value at 1 and worst at value 0.
9. Area Under the Curve (AUC) shows ability of the model to distinguish between FN and FP classes and used as a synopsis of receiver operator characteristic (ROC) curve that narrates about performance of binary classification algorithm. .

Above considerations used to measure the performance of the model are shown in the table 1. Graphical representations of the results are shown in fig 4 and fig 5.

Table 1: Experimental results

Algorithm	Accuracy %	Precision	Recall	F1-score	ROC-AUC score
KNN	69.80	65.17	85.75	74	69.76
LR	56.70	67.24	26.58	38.10	56.78
DT	49.05	48.90	40.55	44.38	49.08
RF	63.18	61.22	74.45	66.97	63.15
AdaBoost	74.39	65.17	85.75	74	69.76

From table AdaBoost Classifier has the maximum accuracy while Decision trees have the least one.

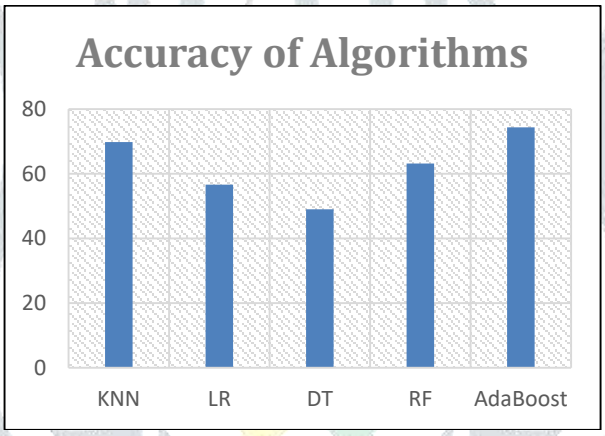

Fig 4: Accuracy of the Algorithms

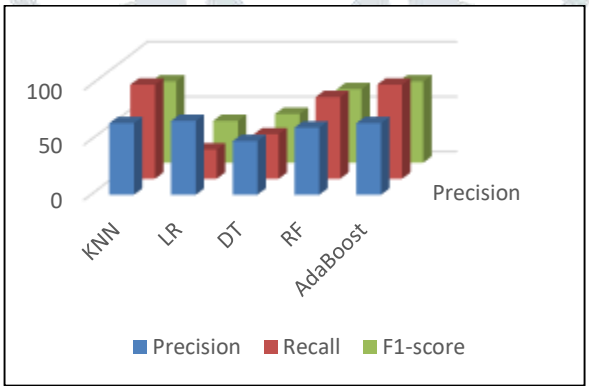

Fig 5: Precision, Recall and F1 scores of the algorithms

IV. Conclusions

This research study sites empirical analysis of selected supervised classification algorithms over Twitter's dataset. Selected machine learning classifier algorithms were implemented in python language. Performance accuracy of the classifiers is evaluated by constructing a confusion matrix. Out of total instances, incorrectly classified and correctly classified instances are taken into considerations through precision, recall, F1 score and ROC-AUC scores. Based on these scores, the accuracy value in terms of percentage is calculated. Final conclusion is drawn as AdaBoost Classifier has the maximum accuracy while Decision trees have the least one over Twitter dataset.

.

V. References

1. Saima Jan, Rahila Ruby , Peer Taha Najeeb , Mudasir Ahmed Muttoo, "Social Network Analysis and Data Mining" by © 2017, IJCSMC Saima Jan et al, International Journal of Computer Science and Mobile Computing, Vol.6 Issue.6, June- 2017, pg. 401-404
2. Boyd, D., & Ellison, N. (2009). Social network sites: Journal of Computer-Mediated Communication,13(1),RetrievedDecember10,2007from:http://jcmc.indiana.edu/vol2013/issue2001/boyd..htl
3. Web resource at https://www.kaggle.com
4. Wasserman, S., & Galaskiewicz, J. (1994). Advances in Social Network Analysis: Research in the Social and Behavioral Sciences. Thousand Oaks: Sage Publications.
5. Jiawei,Han and Micheline Kamber, Data mining: concepts and techniques San Francisco, California, Morgan Kauffmann, 2001.
6. Muthuselvi Et Al, Information Retrieval From Social Network, IJCA Proceedings On E-Governance And Cloud Computing Services – 2012,Vol-4,2012
7. Senol Zafer ERDOGAN Et Al, A Data Mining Application In A Student Database , Journal Of Aeronautics And Space Technologies, Vol-2, Issue-2,2006