Analysis of Data Using Machine Learning Approaches in Social Networks

Fatih Ertam
Informatics Department
Firat University
Elazig, Turkey
fatih.ertam@firat.edu.tr

***Abstract*— The amount of data circulating on the Internet is increasing day by day. With the increasing use of social media in particular, the importance of analyzing these data is increasing. The use of machine learning approaches to analyze large amounts of data is still popular today. Today, the social network Facebook is the most popular social networking sites. In this study, some data taken on Facebook were analyzed by machine learning approaches and compared with performance metrics. Logistic Regression (LR), Random Forest (RF) and Adaboost (AB) were used for machine learning approaches. Performance metrics used for comparison are precision, recall and F1 score. Confusion matrix values and Receiver Operating Characteristic (ROC) curves for the results are also presented. It was observed that the results of the RF and LR studies were close to each other and gave better results than the study done with the AB.**

Keywords*— Machine Learning; Social Media; Big Data; Logistic Regression; Random Forest; Adaboost; Performance Metrics

I. INTRODUCTION

The use of the Internet has become one of today's indispensable activities. Especially the popularity of social networks is one of the most important reasons for the increase of internet usage. According to Hootsuite's "Digital in 2017 Global Overview" report, half of the population in the world is said to be an active Internet user [1]. According to the same report, approximately 2.8 billion people actively use social media. Active social mobile media users are estimated to be about 2.55 billion people. Within social networks, Facebook is ranked first in terms of number of members and usage rate. Facebook founder Mark Zuckerberg announced that in June of 2017, the number of users reached 2 Billion people. Due to the high usage rate, social networking sites have become a commercial area for many firms to use for data analysis [2], [3]. Machine learning approaches are useful for analyzing data to make meaningful results. Analysis of data from social networks has provided the emergence of a new field of social media mining. Thanks to the social media mining, data mining techniques can be used together for the analysis of social network data and good results can be obtained [4]. A data set of Facebook data was used in the study [5, 6]. This data set is given by comparing machine learning approaches with LR, RF and AB algorithms, precision, recall and F1 Score performance metrics.

978-1-5386-0930-9/17/$31.00 ©2017 IEEE

The confusion matrix table for the algorithm results and the ROC curves for the classes are also given.

In the second part of this work, we have provided information about the machine learning approaches used and the performance metrics used. Experimental studies in the third section are presented. In the last part, the results are discussed.

II. MATERIAL AND METHOD

This section provides information on LR, RF and AB from machine learning approaches used to analyze social networking data. To measure the performance of the data mining algorithms used, we show how the preferred performance metrics, precision, recall and F1 Score metrics are calculated.

A. Logistic Regression

LR is a classification approach that is effective for situations where variables do not always consist of quantitative values. It is preferred for binary and multiple classification approaches. The main goal is to determine the probability of obtaining another dependent variable by using independent variables [7]. Whatever the values of the variables in the LR classifier, the result is between 0 and 1. For simplicity, it presents a practical approach, especially when analyzing big data.

B. Random Forest

Although RF was first proposed in 1995, its actual widespread use is after 2001. This algorithm has been developed as a community learning method using decision trees [8]. Each decision tree is constructed by applying a bootstrap sampling of the data. The feature size used for each node is randomly selected from among all features. In order to classify the data, each sample vector is decided on each tree of the forest. It is one of the preferred classifiers for classifying big data because of the simplicity and random choice of variables.

C. Adaboost

AB is a preferred classifier over time, especially when there are a large number of parameters used when deciding. The AB performs the classification process by combining the weights of the feature classifiers created by the use of training data [9] - [11]. Can be used with different classifiers. At the beginning of the training, equal weight is applied to each sample. It is multiplied by the weights given and the responses given by each classifier to all samples. In the end, the least faulty attribute

classifier is chosen. The sample weights are updated by increasing the weights of the samples of the selected attribute classifier. With new weights new performance of each attribute classifier is determined. The main classifier is added by selecting the best performing classifier. This process continues as long as the error rate falls.

D. Performance Metrics

The performance metrics used to analyze the data are the precision, recall, and F1 Score metrics. Precision or positive predictive value is obtained by dividing the true positive (tp) value by the sum of true positive and false positive (fp). Recall or sensitivity is obtained by dividing the true positive value by the sum of true positive and false negative (fn) [12]. The F1 Score metric is the harmonic mean of the precision and recall metrics [13]. The performance values used between Equations 1-3 are shown.

$$precision = \frac{tp}{tp + fp} \tag{1}$$

$$recall = \frac{tp}{tp + fn} \tag{2}$$

$$F1\ score = \frac{2 * tp}{2 * tp + fn + fp} \tag{3}$$

III. Experimental Studies

The attribute information of the data set used is given in Table-1. The data were subjected to the training and testing process according to the 10-fold crossover method.

TABLE I Features and Explanation

No	Explanation
1	Total liked pages
2	Category
3	Month of publication
4	The day the submission was published
5	Delivery time
6	Whether or not the advertisement is paid
7	Number of singles who saw
8	Clicks number
9	Number of singles who click anywhere in the post
10	Number of people clicking anywhere in the post
11	Any number of clicks in the post
12	Total number of people who like the page
13	The number of people who liked the page afterwards
14	Number of people who liked and joined the page
15	Number of posts in the post
16	Likes number
17	Number of shares sent
18	Sum of likes, comments and shares
19	Type

Attribute 19 is used as the class. These classes are status, video, link and photo. The confusion matrix values of the classifiers are shown in Figure 1-3.

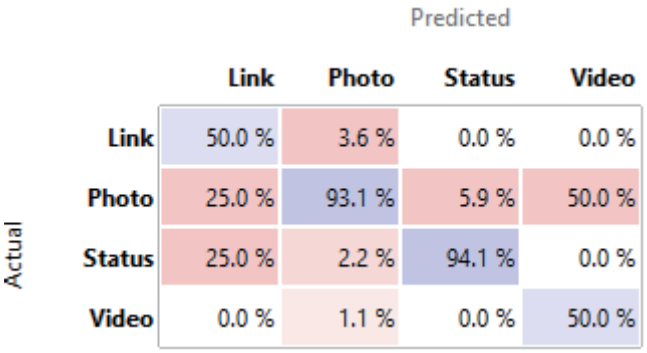

Fig. 1. Confusion matrix for LR

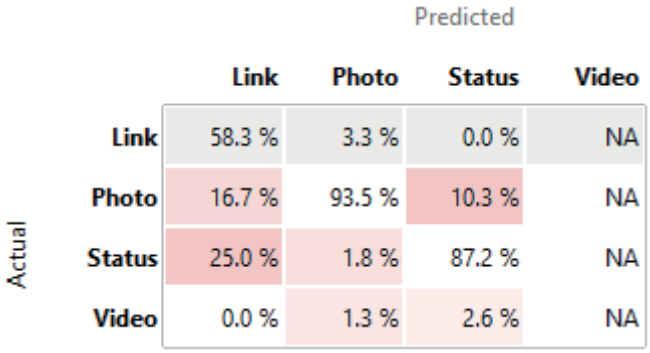

Fig. 2. Confusion matrix for RF

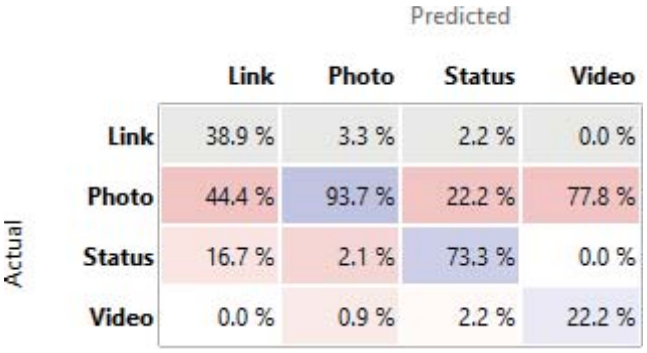

Fig. 3. Confusion matrix for AB

Performance metrics according to confusion matrix values are given in Table-2.

TABLE II Evaluation Results

Classifier	Precision	Recall	F1 Score
RF	0.901	0.922	0.909
LR	0.907	0.918	0.909
AB	0.884	0.886	0.885

According to the measurement results, the best precision value is the classifier's LR. It is seen that the classifier's RF is the best of the recall value. Classifiers with the best F1 Score value are RF and LR.

The ROC curves obtained for each class are shown in Figures 4-7. When the true positive values in the ROC space are high and the false positive values are low, a classifier indicates a good classification. For this reason, it is expected that a good classifier will have values on the upper left of the graph given in Figures 4-7.

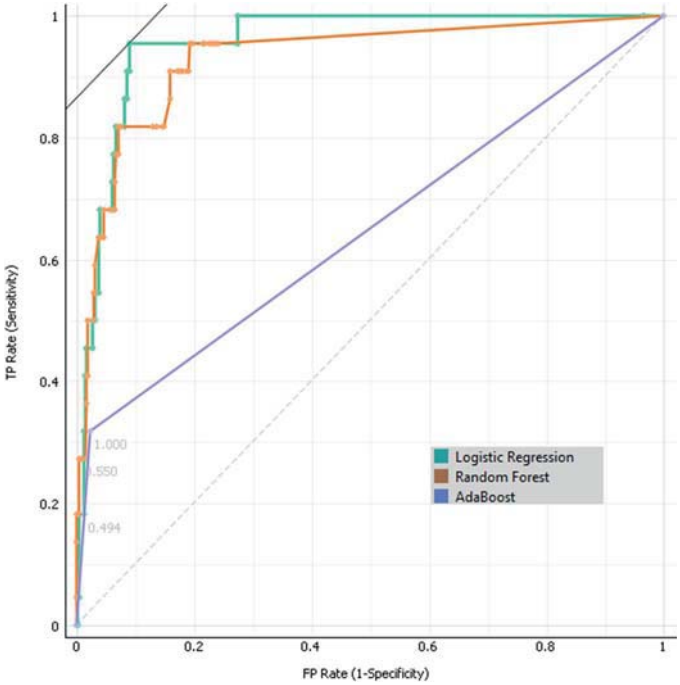

Fig. 4. ROC curve of Link class

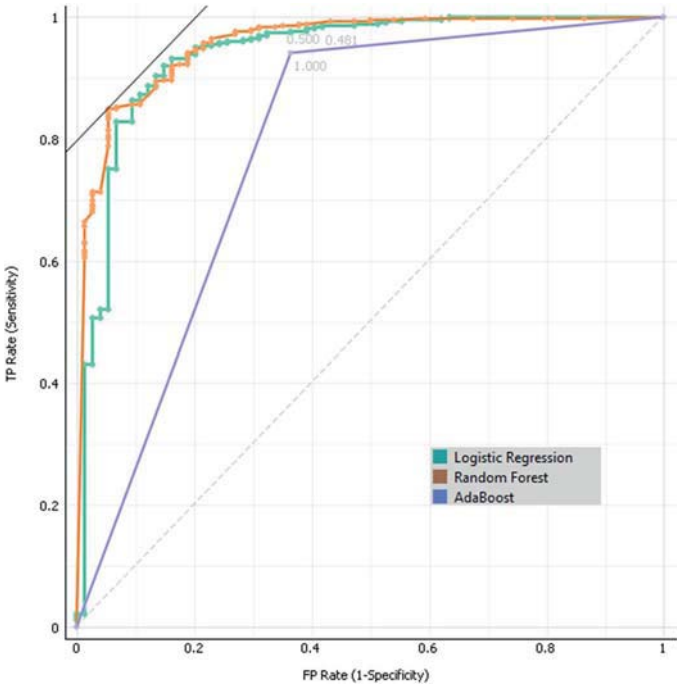

Fig. 5. ROC curve of Photo class

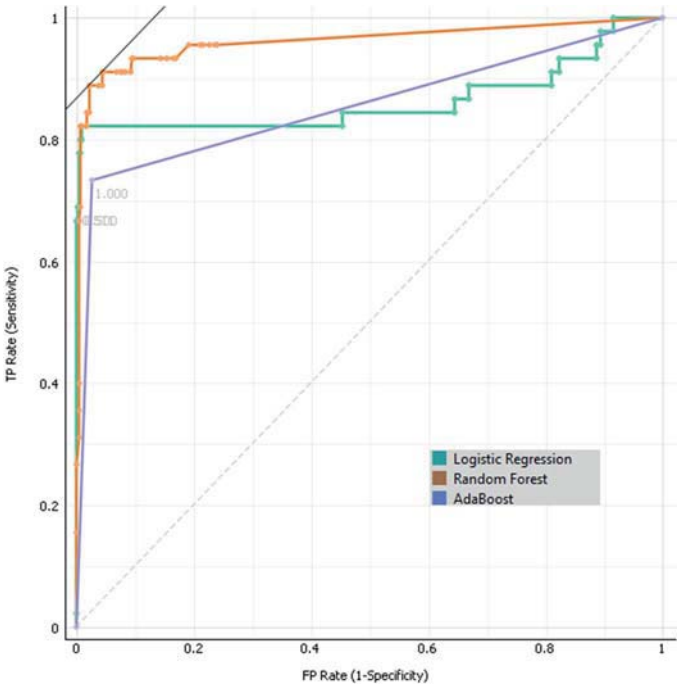

Fig. 6. ROC curve of Status class

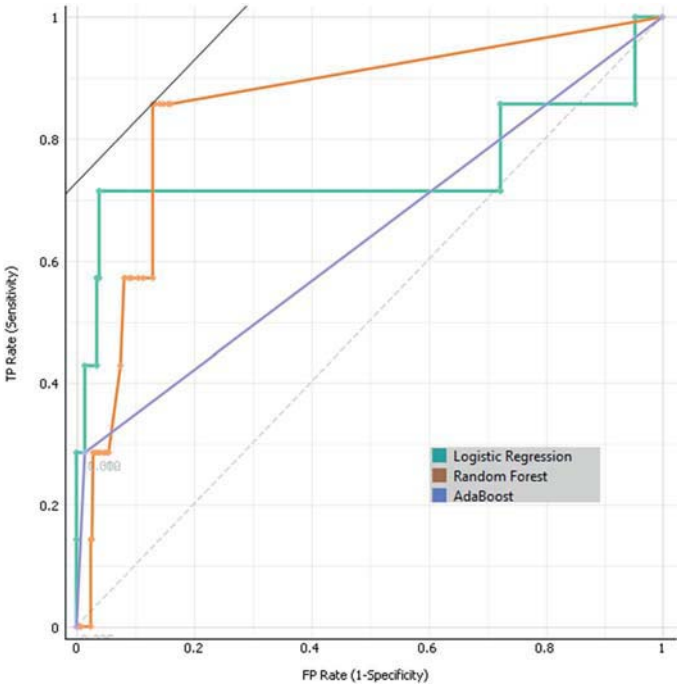

Fig. 7. ROC curve of Video class

IV. Results and Discussion

With the increase in the number of devices connected to the Internet, the amount of data to be analyzed on the internet has reached incredible dimensions. Especially with the spread of social networks, internet use is increasing day by day. Social networks like Facebook make it easy for people to be interested and share data internally. In this study, we have classified data

on Facebook data with data mining algorithms. RF, LR and AB classifiers have been chosen as classifiers. In particular, the performance of these classifiers has been compared because of their rapid performance in classifying big data. Recall, precision and F1 Score metrics were used to compare performance. In order to use these metrics, the confusion matrix table belonging to the classifiers was created and the real and estimated values were determined.18 attributes and 4 classes are defined as data class. Link, Photo, Status and Video classes are used as the class. The data were tested by a 10-fold crossover method. No classifications of the Video class with the RF classifier have been found to be accurate. It has been observed that all classifiers are correctly classified over 90% of the Photo class. The F1 score metric was found to be equal in the RF and LR classifiers, while the AB classifier showed lower output. Performance metrics for the RF and LR classes are close to each other. It has been observed that the results obtained with the AB classifier are worse than the RF and LR classifiers, but there is no significant difference. The ROC curves for each class were generated. When the ROC curve is taken into consideration, the best value for the Link class is observed in the LR class. For other classes it has been observed that the RF classifier has reached its best accuracy.

REFERENCES

[1] S. Kemp, "Digital In 2017: Global Overview," Http://Wearesocial.Com, 2017. [Online]. Available: http://wearesocial.com/sg/blog/2017/01/digital-in-2017-global-overview.

[2] D. Korschun and S. Du, "How virtual corporate social responsibility dialogs generate value: A framework and propositions," J. Bus. Res., vol. 66, no. 9, pp. 1494–1504, 2013.

[3] W. G. Mangold and D. J. Faulds, "Social media: The new hybrid element of the promotion mix," Bus. Horiz., vol. 52, no. 4, pp. 357–365, 2009.

[4] R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining. Cambridge: Cambridge University Press, 2014.

[5] U. M. L. Repository, "Facebook metrics Data Set," www.ics.uci.edu, 2016. .

[6] S. Moro, P. Rita, and B. Vala, "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach," J. Bus. Res., vol. 69, no. 9, pp. 3341–3351, 2016.

[7] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," Journal of Biomedical Informatics, vol. 35, no. 5–6. pp. 352–359, 2002.

[8] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[9] P. L. Bartlett and M. Traskin, "AdaBoost is consistent," J. Mach. Learn. Res., vol. 8, pp. 2347–2368, 2007.

[10] P. Viola and M. Jones, "Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade," in Proceedings of Neural Informatin Processing System, 2001, no. 14, pp. 1311–1318.

[11] G. Rätsch, T. Onoda, and K. R. Müller, "Soft margins for AdaBoost," Mach. Learn., vol. 42, no. 3, pp. 287–320, 2001.

[12] K. Ting, "Precision and Recall," in Encyclopedia of Machine Learning, 2011, p. 1031.

[13] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize F1 measure," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, vol. 8725 LNAI, no. PART 2, pp. 225–239.