iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34626

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Prediction of Online News Popularity using Machine Learning

Prachi[1], Richa Agarwal[2], Saksham Kaushal[3], Ritvik Kaul[4], Saurabh Bisht[5]
[1]*Assistant Professor, Dept. of Computer Science and Engineering, Inderprastha Engineering College, UP, India*
[2, 3, 4, 5]*Student, Dept. of Computer Science and Engineering, Inderprastha Engineering College, Uttar Pradesh, India*

Abstract: ***In today's world, online news is the source of news and information for many people. The popularity of online news can be dependent on various factors like the number of shares by the readers or the number of likes and comments on the news etc. In our research paper, we intend to find the best machine learning algorithm to predict the popularity of the news article. The dataset is taken from the UCI repository. We will apply five different algorithms such as random forest, logistic regression, etc and then compare the performance of these algorithms based on various performance measures such as accuracy and precision. Random forest showed the best results with the highest accuracy. Our work can be used by online publishers to predict the popularity before publication.***

Keywords: Machine learning, Model Selection, Classification

I. INTRODUCTION

Online news is a primary source of information for many people. With the growing use of internet, online news has become the primary source of news and information for a large number of people. The popularity of news depends on various factors like the number of shares, number of likes and, the number of comments on the articles. These factors play an important role to measure the news' popularity. Any news article that is shared more becomes more popular as more people have read that article. The number of shares on any article is an important factor which can be used to determine the popularity of news. The popularity prediction is a binary classification problem as it classifies the news articles as popular and unpopular. This prediction of news popularity can be used by news agencies and publishers to maximize profits and in getting a higher number of advertisements**.** Different machine learning algorithms are planned to implement on the dataset to evaluate and compare their performances using various performance measures.

II. LITERATURE SURVEY

News popularity prediction is a new research topic so there's still plenty of space for further research work in this area. The work done so far in this domain mainly focused on the attributes of online content for estimating future popularity. Frenandes et al.[1] has used the dataset taken from UCI Repository and implemented various machine learning algorithms namely Random Forest, Adaboost, Support vector machine , K-Nearest Neighbors and Naive Bayes. The random forest method performed the best giving an accuracy of 67% and roc of 73%. The accuracy of KNN and Naive Bayes is 67%.The other two methods i.e. SVM and Adaboost, both gave an accuracy of 66%. Here the Random Forest method gave highest accuracy. Hensinger et al. [5] presented that the news choices of each individual is shaped based on various different factors. After reading, people like the news article, share the news, write comments on the news articles. They share their opinion with public and this helps in making the article popular. The click of a user on a news article is influenced by many factors such as articles' position on the web page, timing, topic, number of images, additional media, linguistic style. The method of binary SVM and Ranking SVM is used for the prediction of news articles of 10 different news agencies. Joe Maguire et al. [4] has shown that to predict the popularity of posts on social media accurately and precisely , machine learning can be used. Various machine learning algorithms are used namely Naive Bayes, Perceptron algorithm and linear Support Vector Machine. The Linear SVM method gave the accuracy of 85% which is the highest followed by Naïve Bayes with and accuracy of 81% and Perceptron Algorithm with an accuracy of 73%.Yangjie Y. et al. [3] presented the research which is useful in the design of news recommendation system and personalization of news. An analysis of user behaviour of habits like news viewing, liking, commenting and sharing was done. This analysis was done on the data that was taken from yahoo news which was collected for about 2 month period. The method used is Latent Dirichlet Allocation (LDA) for the analysis. Jiahui L. et al. [2] presented the research which can help in developing the recommendation system of personalized news for an individual on Google news. A log analysis of user click on Google news was conducted which showed the variation in user's interests. Some approaches used the user's comments on the news articles for popularity prediction by analyzing user's comments on different types of news articles.[6]

III. METHODOLOGY

The methodology for the online news prediction is as follows:

A. The dataset was taken from UCI Repository which was released by the popular news website known as Mashable.com. The dataset has 39797 articles and 61 attributes. The data is collected for about 2 years of period, from January 2013 to January 2015.
B. The data is then cleaned and for the data cleaning process, the data was checked for null values and outliers of the data were removed using the Inter quartile range(IQR) method.
C. The mean of the target variable i.e. 'no of shares' is taken and it is considered as the threshold value for the further classification of news as popular or unpopular.
D. The dataset is split into training and testing data in the ratio of 75:25.
E. The machine learning algorithms are then applied to the dataset. Random forest, Logistic Regression, Naive Bayes, K-Nearest Neighbors and, Support vector machine algorithms are the methods which are applied on the dataset.
F. The algorithms are then compared based on three performance measures i.e. accuracy, precision and, ROC.
G. After comparing, the best model among the four methods was selected.

IV. RESULT

The developed system will predict if the news will become popular or not popular. Various machine learning algorithms are applied on this dataset after the data cleaning process and then the best performing algorithm is selected among them using the different performance measures.

S. No	Model	Accuracy	Precision	ROC
1	Logistic regression	0.692	0.693	0.500
2	Random Forest	0.710	0.727	0.571
3	SVM	0.693	0.693	0.500
4	KNN	0.667	0.720	0.551

Table -1: Values after applying algorithms

Table 1 shows the comparison between the four machine learning algorithms namely Random forest, Logistic Regression, Support Vector Machine and K-Nearest Neighbors that are applied on the dataset. Three Performance measures are used to evaluate the performance of these algorithms and compare them. On comparison, it clearly illustrates that the model which performed the best is Random Forest.

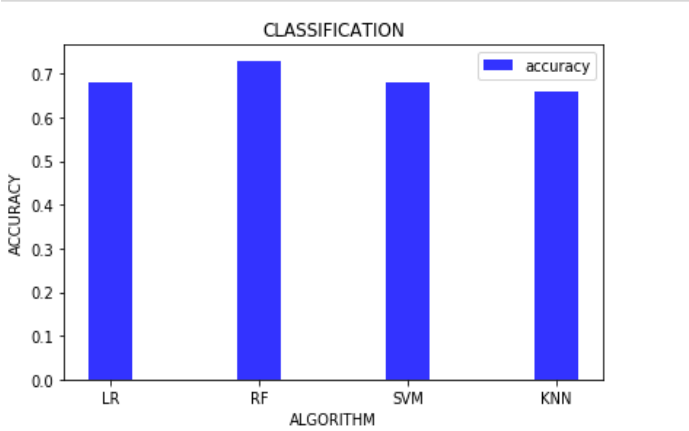


Fig-1: Comparison of various algorithms based on accuracy

Figure 1 showed the comparison of the four algorithms based on accuracy. Random forest method gave the highest accuracy among all.

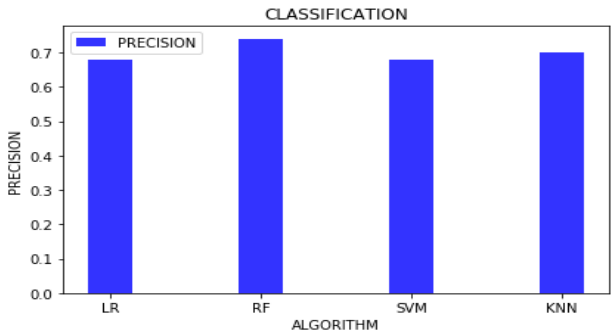


Fig-2: Comparison of various algorithms based on precision

Figure 2 showed the comparison between algorithms based on precision and among all random forest method gave highest precision followed by KNN method.

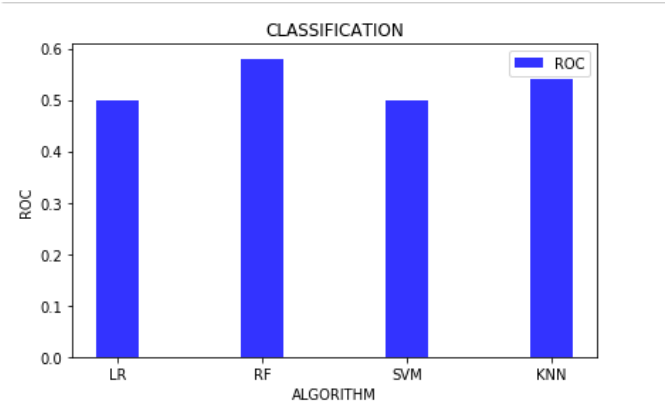


Fig-3: Comparison of various algorithms based on roc

Figure 3 showed the comparison of the algorithms based on roc and Random forest gave the highest roc followed by KNN.

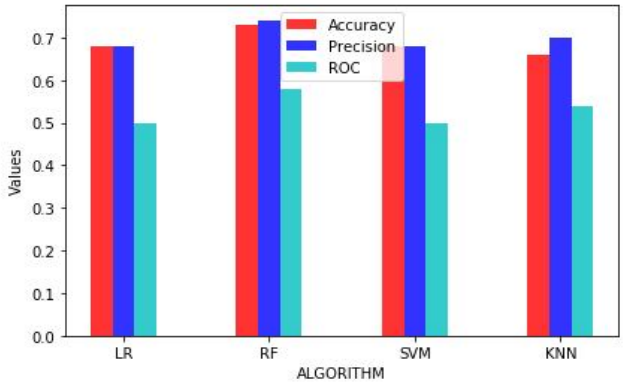


Fig-4: Comparison of various algorithms based on various performance measures

Figure 4 shows comparison among the four algorithms namely Random forest, Logistic Regression, SVM and KNN based on three performance measures accuracy, precision and roc. Among all the algorithms random forest performed the best.
It is observed that in the proposed work, Random Forest has better performance in all evaluation measures. This method gave an accuracy of 0.71 and precision of 0.727.

V. CONCLUSIONS

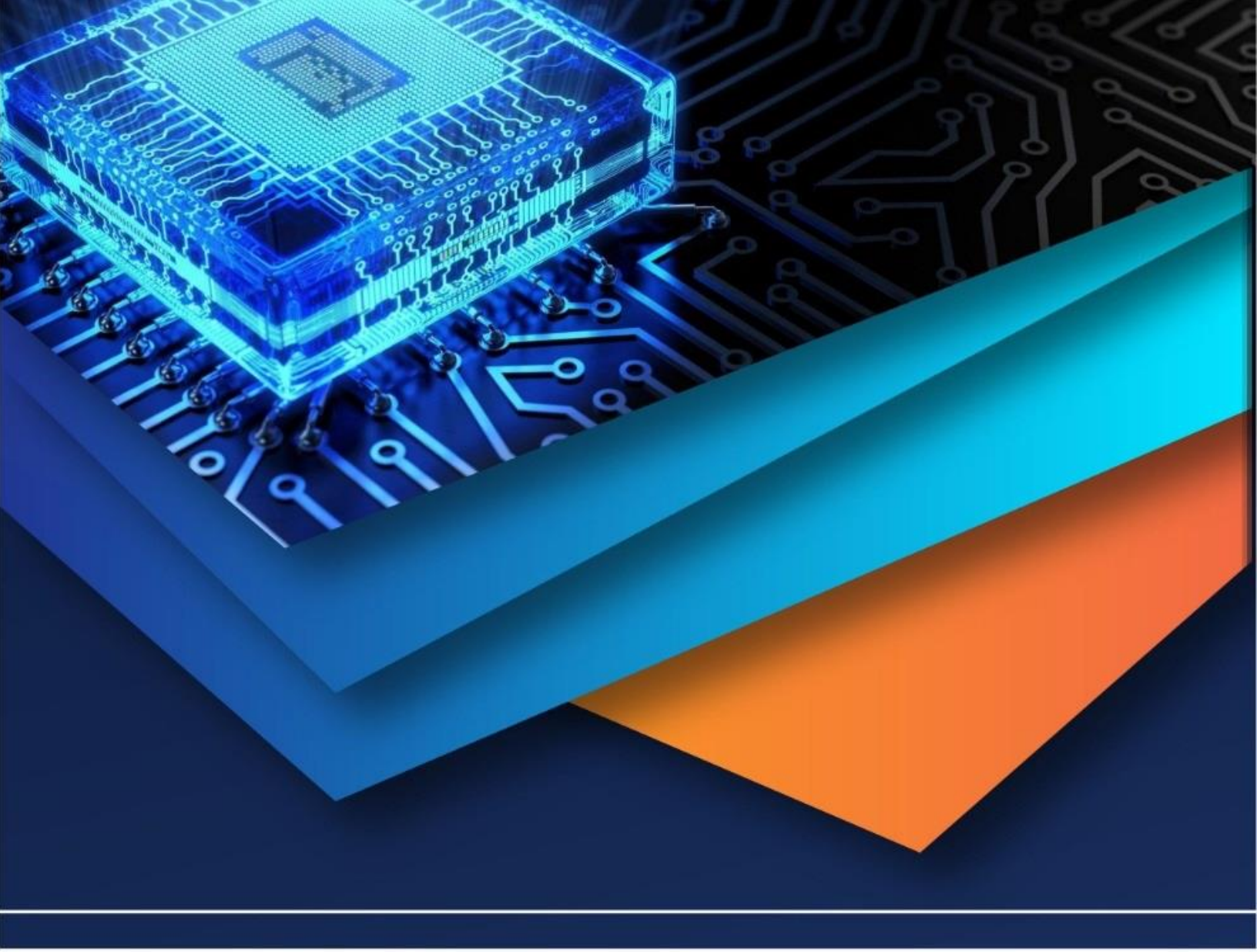
News popularity prediction is a comparatively new research topic and there is plenty of space for further research in this domain. It is still a research area to find out the factors which make particular news a popular one. Here the data is taken from UCI Repository and the dataset split into training and testing ratio of 75:25. The target variable used for prediction is 'number of shares' and 4 machine learning methods are applied on the dataset and the method which performed the best is random forest. Random forest method gave the best accuracy and precision of 0.710% and 0.727% respectively. Logistic regression and Naïve Bayes showed significantly good results for accuracy i.e. 0.692% and 0.693% respectively and K-Nearest Neighbor gave good precision results of 0.72%. Hence it can be seen that the random forest gave the best results among all the implemented algorithms.

VI. ACKNOWLEDGEMENT

The authors would like to extend their esteemed thanks to Ms. Prachi, Professor, CSE Department, Inderprastha Engg. College.

REFERENCES

[1] Kelwin Fernandes, Pedro Vinagre and Paulo Cortez (2015), 'A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News', Springer (EPIA 2015), pp. 535-546.
[2] Jiahui Liu, Peter Dolan and Elin Rønby Pedersen (2010), 'Personalized News Recommendation Based on Click Behavior', Proceedings of the 15th international conference on Intelligent user interfaces (IUI'10), Hong Kong, China, ISBN: 978-1-60558-515-4, pp. 31-40.
[3] Yangjie Yao and Aixin Sun (2013), 'Are Most-viewed News Articles Most-shared?', pp. 1-12.
[4] Joe Maguire and Scott Michelson, 'Predicting the Popularity of Social News Posts', Machine Learning Report, Standford University.
[5] Elena Hensinger, Ilias Flaounas and Nello Cristianini (2012) 'Modelling and predicting news popularity', Pattern Anal Applic, Springer, pp. 623-635.
[6] S. Petrovic A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida, "Predicting the popularity of online articles based on user comments," in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. ACM, 2011, p. 67.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY