DOI: https://doi.org/10.24113/ijoscience.v5i1.181

A Comprehensive Review on Online News Popularity Prediction using Machine Learning Approach

Priyanka Rathord
M.Tech. Scholar
Department of CSE
Radharaman Institute of Technology & Science
Bhopal, M.P, India
parihar28113@gmail.com

Dr. Anurag Jain
Professor
Department of CSE
Radharaman Institute of Technology & Science
Bhopal. M.P, India
anurag.akjain@gmail.com

Chetan Agrawal
Assistant Professor
Department of CSE
Radharaman Institute of Technology & Science
Bhopal, M.P, India
chetan.agrawal12@gmail.com

***Abstract*—With the help of Internet, the online news can be instantly spread around the world. Most of peoples now have the habit of reading and sharing news online, for instance, using social media like Twitter and Facebook. Typically, the news popularity can be indicated by the number of reads, likes or shares. For the online news stake holders such as content providers or advertisers, it's very valuable if the popularity of the news articles can be accurately predicted prior to the publication. Thus, it is interesting and meaningful to use the machine learning techniques to predict the popularity of online news articles. Various works have been done in prediction of online news popularity. Popularity of news depends upon various features like sharing of online news on social media, comments of visitors for news, likes for news articles etc. It is necessary to know what makes one online news article more popular than another article. Unpopular articles need to get optimize for further popularity. In this paper, different methodologies are analyzed which predict the popularity of online news articles. These methodologies are compared, their parameters are considered and improvements are suggested. The proposed methodology describes online news popularity predicting system.**

***Keywords*— News articles, Online news popularity, Popularity prediction**

I. Introduction

Various businesses are keen to know what will be the future demand of online visitors. Accordingly they are ready to take immediate decisions to implement their ideas on the available platform[1]. It is necessary to keep ready their resources with immediate market demand. In this case they get to know if particular article will become popular, and along with its popularity if it will get particular popularity rank then they can put their advertise and increase revenue. Popularity prediction is useful in many applications like media advertising, estimation of movie revenue, traffic management, economic trends forecasting.

Popularity prediction is hard to capture as it depends upon various factors like its topic, text, timing, article's position on the web page, language, similarity with world's event, same subject historical popularity, time from news publication, season of article popularity, relevance to the physical world popular events. Generally news are studied according to binary classification i.e. popular and unpopular[1-5]. Prediction of the popularity is considered into 2 parts i.e. popularity prediction before publication of news and popularity publication after publication of news. Mostly popularity prediction before news publication is considered for study. In the latest years, different types of prediction methods for different types of web information have been proposed. This study focused on prediction of large visitors' attention to particular news articles, its reasons, evaluated methodologies, considered parameters and improved results.

This paper is arranged as follows. Section II reviews related work of news popularity prediction process while section III describes various existing methodologies used in popularity prediction. Section IV gives comparative analysis of various prediction methods. Section V describes proposed methodology and parametric explanation. Section VI gives suggestions for improvements.

II. RELATED WORK

Kelwin Fernandes, Paulo Cortez and Pedro Vinagre in [1] proposed a system for Intelligent Decision Support or called as (IDSS) and focused on predicting whether an article will be popular before getting published, and then used optimization techniques to improve few of the article features so that maximum popularity could be achieved for that article, prior to its publication. They used 47 out of 60 features and using Random

Forest an accuracy of 67% was achieved and optimization was done using Stochastic Hill Climbing. He Ren and Quan Yang in [2] optimized the work done in [1] by making use of Machine Learning techniques including Mutual Information and Fisher Criterion to get maximum accuracy for feature selection, based on which prediction for popularity of news article was done. Using this method, they got an accuracy of 69% using Random Forest using top 20 features.

H. Muqbil, AL-Mutairi, Mohammad Badruddin, Khan in [3] had predicted the popularity of trending Arabic Articles taken from the Arabic Wikipedia based on external stimulus including number of visitors. This paper used Decision Tree and Naïve Bayes for prediction and compared the two models.

Elena Hensinger, Ilias Flaounas and Nello Cristianini in [4] had predicted the popularity of the article, on the basis of the number of views that the article had on the day it was published. It used RSVM method to predict popularity which is done on the basis of the title of the news article, its introductory description, the place and date of its advertisement.

Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, Jérémie Leguay, Arnaud Limbourg and Serge Fdida in [5] talk about how the popularity of a news article can be predicted on the basis of the number of users who commented on that article over a short period (in hrs) of time soon after the article was published.

Roja Bandari_ Sitaram Asury Bernardo Huberman, in [6] predicted popularity on twitter with accuracy of 84% using regression and classification techniques, by considering following attributes—the source which posted the article, category of the news, use of appropriate language and names of people mentioned in the article. Score assignment of each features is done and accuracy was found out using Bagging, J48 Decision Trees, SVM and Naïve Bayes.

In [7], I. Aprakis, B. Cambazoglu and M. Lalmas, do a cold start prediction, where they acquire their data from Yahoo News and predict the popularity. For prediction they use two metrics: Number of times the article was shared or posted on twitter and the number of views for that article.

Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, Matt Stempeck in [8] present a qualitative as well as a quantitative study of the longevity of online news article by analyzing the strong relation between reaction on social media and visits over time. This study shows that both are equally important for understanding the difference among classes of articles and for predicting future patterns.

Gabor Szabo, Bernardo A. Huberman in [9] predicts the longterm popularity of online articles which were taken from YouTube and Digg by analyzing the views and votes for these articles.

Zhiyi Tan et al. [24] treated the popularity of online videos as time series over the given periods and propose a novel time series model for popularity prediction. The proposed model is based on the correlation between early and future popularity series. Experimental result on real world data have demonstrated that the proposed model outperforms several existing popularity prediction models.

Jong Gun Lee et al. [25] predicts popularity of online content based on features which can be seen by an external user, including number of comments and the number of links in the first hours after the content publication. This work can predict lifetime based on number of threads (5–6 days) and number of user comments (2–3 days). It is an optimized paper for [23] using survival analysis.

Swati Choudhary [26] used genetic algorithm to get the optimum attributes and further classified the data using different classifiers and obtained the highest accuracy of 91.96% with naïve bayes classifier.

In this project, we will use machine learning techniques to solve a multi-classification problem, which is to predict if an online news article will become popular or unpopular or average prior to publication. The popularity is characterized by the number of shares. If the number of shares is higher than a pre-defined threshold, the article is labeled as popular, otherwise it is labeled as unpopular. Thus, the problem is to utilize a list of article's features and find the best machine learning model to accurately classify the target label (popular/unpopular) of the articles to be published.

Table 1: Comparative analysis of various popularity prediction methods

Method	Description	Conclusion
Random Forest [1]	Training data was 70% and validation sets was 30% by using random holdout split. No. of trees{10,20,50 ,100,200,400}	The best obtained result AUC = 0.73 is 23 percentage points higher than random classifier. Best model for prediction.

SVM [4]	No. of features {1,5,10,20}	Average accuracies were above 60% in 8 out of 10 cases. For new York times and Seattle times above 70%
AdaBoost [1]	No. of tress {10,20,50,100,200,400}	AUC = 72%
KNN [1]	No. of neighbors {1,3,5,10,20}	AUC = 67%
Naïve Bayes [1]	Metrics computed over the union of all 29 test sets.	AUC = 65%
Linear Regression [2]	Target values were discretized to binary categories	Accuracy = 66%
Logistic Regression [11]	Data is classified and stochastic gradient ascent rule was used to implement it.	Accuracy = 66%
Genetic Algorithm [12]	Used genetic algorithm to get the optimum attributes and further classified the data using naïve bayes.	Accuracy =91.96%

III. Dataset & Feature Selection

A. Data collection

Our dataset is provided by UCI machine learning repository [1], originally acquired and preprocessed by K.Fernandes et al. It extracts 59 attributes (as numerical values) describing different aspects of each article, from a total of 39644 articles published in the last two years from Mashable website.

B. Feature selection

It is also known as variable selection, attribute or subset of variables, is used in machine learning or statistics to select a subset of features to construct data description models. Two important aspects of selection of characteristics are: (i) minimum redundancy and (ii) maximum relevance.

i. Principle Component Analysis (PCA)

The PCA is a commonly used dimensionality reduction algorithm that could give us a less dimensional approximation to the original dataset, while preserving the possible variability. It is a type of pattern recognition in the data. PCA is a powerful tool for data analysis.

Once the data profiles and the compressed data are found, the number of dimensions can be reduced without much information loss.

Steps in PCA are :

- Get some data
- Subtract the mean
- Calculate the covariance matrix
- Calculate the eigen vectors and eigen values of the covariance matrix
- Choosing the components and forming a feature vector
- Derive the new dataset.

ii. Linear Discriminant Analysis

LDA is used to perform dimension reduction in the pre-processing phase for model classification and machine learning applications. The aim of this method is to project a set of data into a small space with good class separability to avoid over-adaptation and also to reduce the cost of the calculation. LDA addresses the situation in which the frequencies within the class are unequal and their performance has been examined on randomly generated test data. The LDA method maximizes the relationship between the class variance and the class variance in a given record. Therefore it guarantees maximum separation.

The general steps of LDA are listed as below:

- Compute the d-dimensional mean vectors for different classes from the dataset.
- Compute scatter matrices (in between class and within class scatter matrix)
- Compute eigen vectors (e_1, e_2, …, e_d) and corresponding eigen values (λ_1, λ_2, …, λ_d) for the scatter matrices.
- Sort the eigenvectors by decreasing eigen values and choose k eigen vectors with the largest eigen values in order to form a d x k dimensional matrix - w (where every column represents eigenvector).
- Use this d x k eigen vector matrix to transform samples in to the new subspace. This can be summarized by the matrix multiplication: Y = X x W (where X is a n x d – dimensional matrix representing the n samples and y are the transformed n x k – dimensional sample in the new subspace)

iii. Mutual Information

For ranking features, Mutual Information, MI (x,y), is calculated between class values and features. MI can be represented as:

$$MI(x_i, y_i) = KL(p(x_i, y_i)||p(x_i)p(y_i)$$

where p(x, y) =joint probability distribution function of X and Y

p(x) = the marginal probability distribution functions for X

p(y) = the marginal probability distribution functions for Y.

A measure of mutual information is used to calculate the gain of information between characteristics and between characteristics and class attributes.

iv. Fisher Criterion

Another way for ranking efficient features is fisher criterion. For that fish score for every mentioned features is found by:

$$F(j) = \frac{(x_j^{-1} - x_j^{-2})}{(s_j^1)^2 + (s_j^2)^2}$$

Where,

$$(s_j^k)^2 = \sum_{x \in K} (x_j - x_j^{-k^2})$$

The numerator indicates the discrimination between two class, and the denominator indicates the scatter within each class. The larger the F-score is, the more likely this feature is more discriminative.

IV. PREDICTION METHODOLOGIES

A. Random Forest

Random forest is found as best model for prediction. It is learning method for classification, regression. Multiple decision trees are constructed at training time and outputting the classes or prediction. Random forest applies bootstrap aggregation technique which decorrelates the trees by showing them different training sets. For each tree, a subset of all the features can be used. As the number of decision tree increases, the variance of the model can be greatly lowered and Accuracy increases. In Random Forest, 2 main parameters are considered i.e. number of trees and number of features they select at each decision point. Accuracy of prediction increases as more number of trees making decisions. RF improves prediction accuracy as compared to single trees. RF handles larger numbers of predictors and it is faster to predict. RF found to overfit for some datasets with noisy classification tasks. Large number of trees may make the algorithm slow for real-time prediction [1,5].

B. Adaptive Boosting

It is one of the boosting algorithms which combine weak rules into a single strong prediction rule. This algorithm pays higher focus on examples which are misclassified or have higher errors by preceding weak rules. Predictive quality is boosted. This is fast algorithm and no prior knowledge needed about weak learner. But too weak classifier can lead to low margins and overfitting. It is vulnerable to uniform noise. Decision stamps are used with AdaBoost [1].

C. Support Vector Machine

SVM outperform other conventional learning methods for text classification task. SVM method has been used for binary popularity classification. The maximal margin classifier distinguishes 2 data classes with hyperplane. The linear function is described as (w,x) + b. Kernel trick in SVM can be used to improve performance for separating classes which are non-separable with a linear hyperplane. Popular kernels are Gaussian, Polynomial, and Sigmoid kernel. 5 different values for the parameter C have been tested. Best performance per outlet is reported based on the optimal parameter C [3]. Different SVM kernels are used because linear kernel has high bias problem. Polynomial and Gaussian kernels are operational in a high-dimensional, implicit feature space and which are without computing the coordinates of the data in that space. Here, more flexible decision boundaries can be offered. SVM is useful when the data is not regularly distributed or have an unknown distribution. It produces very accurate classifiers and it is robust to noise. SVM has lack of transparency in results and it is computationally expensive, runs slow. As SVM is binary classifier so to do multiclass classification, one class against all others are used i.e. called as pair-wise classifications [5].

D. K-Nearest Neighborhood

This algorithm identifies the k nearest neighbors of 'c', regardless of labels. The input consists of K-closest training examples and output is class membership. Prediction for test data is done on the basis of its neighbor. In case of big data samples, k-NN finds complexity in searching the nearest neighbors for each sample [1].

E. Naïve Bayes

This algorithm is suited when the dimensionality of the inputs is high. Probability is computed for all features and shows the output with highest probability. Algorithm requires small amount of training data to estimate the parameters. Disadvantage if loss of accuracy [1].

F. Linear Regression on Logarithmic Scale (linear log)

Szabo and Huberman proposed this method to predict the popularity of online content by linear regression method. Popularity of YouTube videos and Digg stories was analyzed. A high linear correlation was observed between the log-transformed early and future popularities of online content up to a normally distributed noise. They proposed a simple linear popularity prediction model. Predictions were more accurate. The goal of prediction method is to estimate the number of comments after an article is published [2].

G. Linear Regression

It is the commonly used predictive analysis method. Regression estimations are used to describe the relationship between one dependent variable and one or more independent variable. It predicts trends and future values. It consists of more than just fitting a linear line within a cloud of data points like analyzing and correlation of the data, model estimation and usefulness evaluation. 66% accuracy has shown which was quite desirable [5].

H. Logistic Regression

This regression is conducted when the dependent variable is dichotomous (binary). Data is classified and stochastic gradient ascent rule is used to implement it. Here, logistic regression gave 51.6% accuracy and only focus was given on predicting whether news would be popular or not [5].

V. EVALUATIONS PARAMETERS

Basically 3 modules are followed like data extraction, popularity prediction and optimization. First module is responsible for collecting online articles and their features. While extracting number of words, non-stop words, unique non-stop words, number of links, number of images, videos, category, number of shares etc., the process follows text preprocessing. Various filter methods are available. Among those filter methods mutual information and fisher criterion [5-10] are used. For classification task, machine learning methods have been used which determines the popularity and non popularity of specific news article. If it goes beyond certain threshold then the news becomes popular otherwise it got optimized and decision was given for changes in article.

For optimization purpose stochastic hill climbing search method [1] is used. Here it shows that the without keywords optimization is an easier task as compared to with keywords search.

To predict popularity of online news various classification methods have been used. The metrics which are computed are accuracy, precision, recall, F1, AUC (area under curve).

a) Accuracy

It measures how often classifier makes the correct prediction. It is the ratio of number of correct predictions to the total number of predictions (number of test data points)

Accuracy = correct/predictions

b) Precision

It measures and answers the question: out of the items which are predicted true, how many are actually true?

Precision = tp / (tp + fp)

c) Recall

It answers the question: out of all the items which are true, how many are found to be true by classifier?

Recall = tp / (tp + fp)

d) F-scores

F1-score combines both precision and recall. This score comes in between 0 and 1. 0 is worst and 1 is ideal.

F1 = 2 * [(precision * recall) / (precision + recall)]

e) ROC curve (Receiver Operating Curve)

The curve is plotted as the true positive rate (TPR) against false positive rate (FPR) for various threshold settings. A good ROC curve has a lot of space under it.

f) AUC (Area Under Curve)

Here, the curve is ROC curve. AUC is computed using binned histogram.

VI. PROPOSED PARAMETRIC APPROACH

The purpose of these methodologies is to classify the give data and predict popularity. In literature, Random forest algorithm performs the best. News articles are collected from news website[7]. These are the articles which need to be predicted for popularity. In text preprocessing, the extracted data is converted into suitable format for learning model. In feature extraction process, number of words in title, article is studied. Number of links, images, videos, keywords is extracted. Article category is found out. NLP features are also studied [8-10].

In prediction module, preprocessed data is collected. This data gets separated into training and test sets. Then classification models are applied. The model which gives best result is stored and is used to estimate popularity of an article.

In optimization module, news article's content characteristics are searched. Decision is made and provided to the user which suggests the list of possible changes in article. User can also take decision of her/his own after getting suggestion's list. It may increase the predicted popularity of existing article. Best classification model can be used which has performed well in prediction.

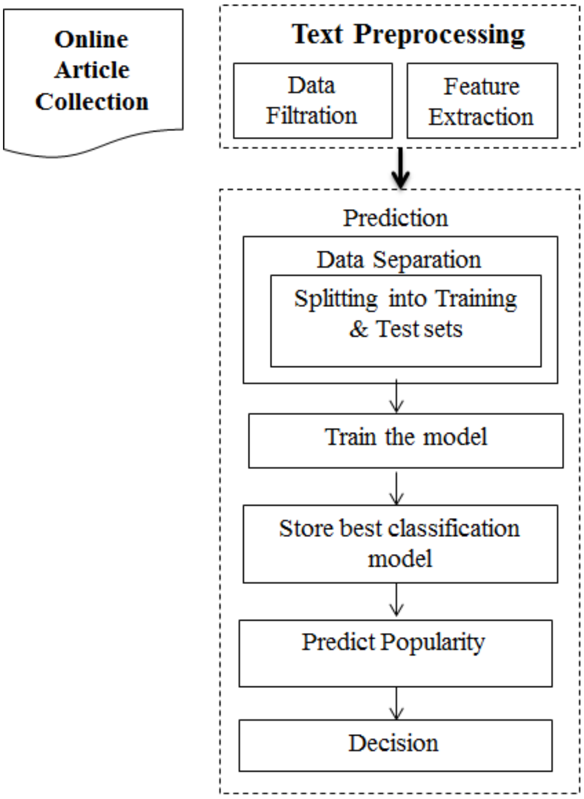

Figure 1: Block diagram of News Popularity Prediction System

More advanced features can be explored. After publication of news article and within specific time period, popularity prediction can be estimated. Different news outlets can be studied for particular news' subject and their predictions can get compared. Factors and features may be analyzed and improvement will give benefit for various service and business industries for their product advertise launching.

Random forest has performed well and given good results. Number of trees can be increased for different set of features. It improves the performance of classification model by improving accuracy result.

VII. CONCLUSION

The analytical study discusses various algorithms used in the process of popularity prediction of news articles. Various news outlets have been considered. How their Features' values are computed, classified into categories and optimized have been studied. The classification models are compared for parameters like accuracy, precision, recall, f1, and AUC. Best result shown by Random Forest classification algorithms. There is scope to improve the classification accuracy and optimization for improving popularity. The popularity can be predicted on number of news' shares, number of likes, number of comments or tweets, number of views etc. Tweet hierarchy also considered on tweeter for news popularity. To improve the value of accuracy, neural network algorithms will be studied and will compare.

REFERENCES

[1] Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News" EPIA 2015, pp. 535–546, 2015.

[2] He Ren, Quan Yang, Stanford University, "Predicting and Evaluating the Popularity of Online News", Machine Learning Project Work Report, 2015, pp. 1-5.

[3] Al-Mutairi, Hanadi Muqbil, and Mohammad Badruddin Khan. "Predicting the Popularity of Trending Arabic Wikipedia Articles Based on External Stimulants Using Data/Text Mining Techniques." Cloud Computing (ICCC), 2015 International Conference on. IEEE, 2015.

[4] Elena Hensinger, Ilias Flaounas, Nello Cristianini, "Modelling and predicting news popularity" Springer, Pattern Anal Applic, 2013, pp. 623–635.

[5] Tatar Alexandru, et al. "Predicting the popularity of online articles based on user comments." Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

[6] Bandari Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity." arXiv preprint arXiv:1202.0332 (2012).

[7] Arapakis, Ioannis, B. Barla Cambazoglu, and Mounia Lalmas. "On the feasibility of predicting news popularity at cold start." International Conference on Social Informatics. Springer International Publishing, 2014.

[8] Castillo, Carlos, et al. "Characterizing the life cycle of online news stories using social media reactions." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014.

[9] Szabo Gabor, and Bernardo A. Huberman. "Predicting the popularity of online content." Communications of the ACM 53.8 (2010): 80–88.

[10] Lee Jong Gun, Sue Moon, and Kave Salamatian. "An approach to model and predict the popularity of online contents with explanatory factors." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.

[11] Bo Wu, Haiying Shen, "Analyzing and Predicting News Popularity on Twitter", International Journal of Information Management, Elsevier, 21st July 2015, pp. 702-711.

[12] Zhiyi Tan, Yanfeng Wang, Ya Zhang, and Jun Zhou, "A Novel Time Series Approach for Predicting the Long-Term Popularity of Online Videos", IEEE Transactions on Broadcasting, 2016.

[13] Lee Jong Gun, Sue Moon, and Kave Salamatian. "An approach to model and predict the popularity of online contents with explanatory factors." Web Intelligence

and Intelligent Agent Technology (WI-IAT), IEEE, 2010.

[14] Swati Choudhary, Angkirat Singh Sandhu and Tribikram Pradhan, “Genetic Algorithm Based Correlation Enhanced Prediction of Online News Popularity” Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing, Springer, 2017, pp.133-144.