INTERNATIONAL STANDARD SERIAL NUMBER
ISSN: 2455-0108

SJIF IMPACT FACTOR: 3.462

IJO-Scince

INTERNATIONAL JOURNAL ONLINE OF SCIENCE

Volume 4, Issue 9, September 2018

www.ijoscience.com

A Study on Online News Popularity Prediction Techniques

Shivangi Bhargava[1], Dr. Shivnath Ghosh[2]

P.G. Student, Department of Computer Science and Engineering, MPCT, Gwalior, India[1]

Associate Professor, Department of Computer Science and Engineering, MPCT, Gwalior, India[2]

***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. With the expansion of the Internet, more and more people enjoy reading and sharing online news articles. The number of shares under a news article indicates how popular the news is. In this project, we intend to study different and the best model and set of features to predict the popularity of online news, using machine learning techniques. The data comes from Mashable, a well-known online news website. 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. Feature selection methods are used to improve performance and reduce features So, it is necessary to know what makes one online news article more popular than another article.**

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

I. Introduction

Daily news paper is the most vital part in everybody's life. People make their opinion after reading the news. They may share their opinions with public like friends, colleagues, family members. It helps to explore their opinions with others and also increase their knowledge. People share their views regarding particular news. The news may relate with day to day life event such as political scenario, education, economy, new opportunities in market, movie release, festival etc. Some people like others opinions and few dislikes it. It depends upon everybody's views and experiences [1]. It may form debates. Now a day scenario is changed. Due to improvement in web technology, single click of user converts and make changes in requirements. In the digital world, online news is primary source of information. News is shared in large numbers through online social networks which usually links to news website. It's become easier to read news on social media and news web sites. News comments are viewable to huge public of world instead of limited group members.

News popularity depends upon various factors like its linguistic style, relevance to the current event, author's image in public, channel history of news etc. After reading, people share the news article, write comments on it, like the news etc. They are sharing their opinions with public. It helps to get publicity for news. The users click on any articles or do not click; it is influenced by many factors such as articles position on the web page, timing, topic, text, additional media [2]. These factors play important role to measure the news‟ popularity. If a news article got maximum number of shares then that news become most readable news. This case is similar for number of comments and number of likes. Now what are the factors which make particular news popular is the research area. This research also helps to optimize unpopular news to become popular or popular news to become more popular. Popularity and unpopularity prediction is binary classification task. Different metrics are used to quantitatively generate popularity. These metrics are giving different levels of user involvement and providing valuable information. News rating improves publication quality. News comments give information of time spent on news page. News sharing creates a good notoriety [3-5]. In this context, relevance of these metrics with each other provides information of what the popularity of news content actually means.

If news became useful and popular among the people then business opportunities grow for marketing companies and also for content providers. Then it becomes valuable asset on the Internet because attention concentrates only on few publications. In the world, companies spend 30% from their budget on online marketing [5]. Early detection of news popularity becomes the next rising star on the internet because these popular news pages can maximize revenues through better advertise launching and placement. The large numbers of prediction methods for various types of web content are proposed in the research of latest years.

Businesses are interested in 2 basic things such as prediction and optimization. Prediction is for knowing what will happen next and optimization is for making the good decision under critical situation such as risk and uncertainty.

A. *Intelligent decision support system Techniques*

Intelligent decision support system (IDSS) has been proposed which analyzes online news before its publication. It predicts if an article will become popular. Online news' popularity is measured by considering communication between web and social networks with factors like number of shares, likes and comments. The popularity of candidate articles is first estimated and changes in unpopular news are suggested in optimization module [1]. Different Techniques for IDSS are discussed below:

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].

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].

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].

K-Nearest Neighbor

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].

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].

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].

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. [13] 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. [14] 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 [15] 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.

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%
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%
SVM-RF [11]	Multiclassification is performed using correlation based feature selection.	Accuracy =73%
Logistic Regression [12]	Data is classified and stochastic gradient ascent rule was used to implement it.	Accuracy = 66%

Genetic Algorithm [15]	Used genetic algorithm to get the optimum attributes and further classified the data using naïve bayes.	Accuracy =91.96%

III. Dataset & Proposed Algorithm

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.

Algorithm : Finding Popular Articles

1: procedure POPULARITY(shares)

2: sum ← 0

3: for each i in shares do

4: sum ← sum + i

5: end for

6: avg ← $\frac{sum}{\text{length(shares)}}$

7: for each i in shares do

8: if i<=median(i)

9: popularity = 0; // popularity = unpopular

10: else if i>= average(i)

11: popularity = 2; // popularity = popular

12: else

13: popularity = 1; // popularity = average

14: end if

15: end if

16: end for

17: end procedure

The purpose of these methodologies is to classify the give data and predict popularity. In literature, machine learning algorithms 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[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.

IV. 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.

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.

V. 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. In this article we reviewed the current state-of-the-art on web content popularity prediction methods. We presented the different prediction methods, reported their performance, and suggested several applications that can benefit from these findings. Even if research on predicting the popularity of web content has been an active area in the latest years there are many avenues that wait to be explored.

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] Kathal, A., & Namdev, M, "Correlation Enhanced Machine Learning Approach based Online News Popularity Prediction", Vol 4, No 3, 2018. Available at: http://ijoscience.com/ojsscience/index.php/ojsscience/article/view/124/101.

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

[13] 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.

[14] 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.

[15] 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.