ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Online News Popularity Prediction using Deep Neural Network with Reduced Features

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: News popularity is the maximum growth of attention given for particular news article. The popularity of online news depends on various factors such as the number of social media, the number of visitor comments, the number of Likes, etc. It is therefore necessary to build an automatic decision support system to predict the popularity of the news as it will help in business intelligence too. The work presented in this study aims to find the best model to predict the popularity of online news using deep neural network.In this work news are predicted into three categories, i.e. popular, average and unpopular. For classification of reduced feature dataset deep neural network, support vector machine and k-nearest neighbor is used. From the experimental results, it is observed that deep neural network outperforms best with respect to support vector machine and k-NN and achieved accuracy of about 76%.

KEYWORDS: Machine Learning, Classification, Popularity Prediction, Correlation Co-efficient, Accuracy.

I. INTRODUCTION

The prediction of the popularity of online news content has remarkable practical values in many fields. For example, by utilizing the advantages of popularity prediction, news organization [1] can gain a better understanding of different types of online news consumption of users. As a result, the news organization can deliver more relevant and engaging content in a proactive manner as well as the organization can allocate resources more wisely to develop stories over their life cycle.

Furthermore, prediction of news content is also beneficial for trend forecasting, understanding the collective human behavior, advertisers to propose more profitable monetization techniques, and readers to filter the huge amount of information quickly and efficiently [1] [3].

The notion of popularity is often expressed by investigating the number of interactions in the web and social networks, for example, click-through rate, number of shares, likes, and retweets.

Tatar et al. [4] demonstrated two types of popularity prediction techniques that are:

- **After publication technique:**More common technique, which uses features capturing the attention that one content receives after its publication. Higher prediction results are expected in after publication technique since utilization of information about the received attention makes the prediction task easier [1] [5] [6] [7].
- **Before publication technique:**Relatively challenging and effective technique. This technique uses only content metadata features that are known prior to the publication of contents instead of using features related to the attention that one content receives after contents release. Although the expected prediction accuracy is comparatively low in before publication method as we are using only metadata features rather than original news content [8], the prediction is more desirable as far as it fosters the possibility of decision making to customize the content before the release of content [9]. In this work, we model popularity prediction problem in before publication technique.

Although popularity prediction of web content has tremendous impacts in many areas, popularity prediction task still faces a bunch of major challenges [4] [9]. First, different factors make prediction difficult, for example, the quality of content or relevance of content to users can influence contents popularity. Second, the relationship between events in the real world and content itself are not only difficult to capture but also hard to further feed into the prediction engine. Third, prediction of complex social interactions and information cascades at the microscopic level are extremely

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

challenging. Fourth, the prediction might also be difficult because of the inaccessible content like context outside the web, local and geographical conditions, and situations which influence the population. Last but not least, the prediction may also be hard based on the network properties e.g. the structure of the networks, and the interplay between different layers of the web.

Previously, researchers try to estimate the popularity by predicting whether or not someone share the news. However, this approach is less informative since we can only identify users share the news rather than how many users share the news. Hence, this paper proposes an extension to the previous popularity prediction models by predicting the number of shares of news using deep neural network in before publication setting.

II. RELATED WORK

Roja Bandari_ Sitaram Asury Bernardo Huberman, in [3] 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 [4], 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.

Joe_Maguire et al. in [12] has shown that the machine learning can be used to accurately predict the popularity of social media post. The model predicted the post's popularity by 85% accuracy. Training data has been collected from Hacker News i.e. from http://news.ycombinator.com/. 4000 posts are acquired. Feature set included the features like the words in the post, the domain the post links to, the time of day the post was created. Naïve Bayes achieved 81% accuracy. The Perceptron algorithm achieved 73% accuracy. The Linear SVM algorithm achieved 85% accuracy on the data.

Tim_Paek et al. [15] studied how people judge the importance of their newsfeed. Authors conducted a study where facebook users were asked to rate the importance for their newsfeed posts and their friends. The classifiers of newsfeed and friend importance have learned to identify predictive sets of features related to social media properties, the message text, and information of shared background. The best performing model achieved 85% accuracy and 25% error reduction for classifying friend importance. The model used to classify newsfeed posts and it achieved 64% accuracy and 27% error reduction. Authors used SVM classifiers using various combinations of features.

Swati Choudhary [10] 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.

1. Proposed Methodology

In the current scenario, an algorithm is proposed which provide a way to predict whether an article will become popular or not. Figure 1 shows the overall architecture for prediction of popularity of online published news. The proposed work is presented for two cases, i.e. case I(for binary classification) and Case II (for multi-classification). The proposed work is designed for optimized feature selection for online news popularity prediction process.

Following diagram describes flow of News Popularity Prediction System:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

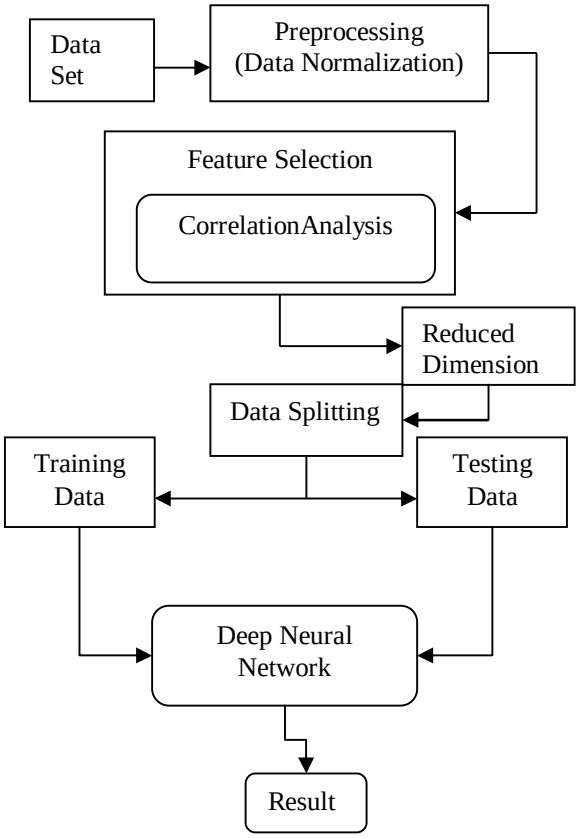

Figure 1: Flowchart of Proposed Methodology

The proposed algorithm works in four stages as discussed below:
Data Preprocessing: In this stage dataset of online news popularity prediction is normalized for further reduction of execution complexity.
Feature Reduction: In this stage the features are reduced using co-relation algorithm.
A bivariate analysis used for measuring the degree of association amongst two vectors say A and B is known as Correlation.In this paper for feature selection Correlation Analysis is performed using Pearson, Spearman and Kendall coefficients which are explained below.
Pearson Correlation Analysis
Pearson correlation coefficient ρ is calculated by the formula as given below:

$$\rho = \frac{E[AD] - E[A]E[D]}{\sqrt{E[A^2] - (E[A])^2}\sqrt{E[D^2] - (E[D])^2}}$$

where:
A stands for the Attribute Vector
D stands for the Decision Vector
E[A] stands for the sum of the elements in A

Spearman Correlation Analysis
Spearman Correlation coefficient σ is calculated by the formula mentioned below:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

$$\sigma = 1-(6\Sigma d_i^2)/n(n^2 - 1)$$

Where,
d_i stands for the difference between the ranks of variables P and Q
n stands for the sample size

Kendall Correlation Analysis

Kendall Correlation coefficient τ is calculated by the formula as given below:

$$\tau = (n_c - n_d)/(1/2n(n - 1))$$

Where,
d_i stands for the difference between the ranks of variables P and Q
n stands for the sample size

After doing Pearson Correlation, Spearman Correlation and Kendall-rank Correlation a list of attributes areobtained that satisfy the respective correlation criteria. After obtaining the three individual results which reduces the number of features using Algorithm 4 discussed below:

Attribute Selection after Correlation

```
procedure ATTRIBUTESELCTION(Dataset)
rows ← nrows(Dataset)
cols ← ncols(Dataset)
pearsonVector ← pearson(Dataset)
spearmanVector ← spearman(Dataset)
kendallVector ← kendall(Dataset)
for each i in 1:cols do
if pearsonVector[i]>0 AND spearmanVector[i]>0
AND kendallVector[i]>0 then
Selection ← true
else
Selection ← false
end if
end for
return dataset[,Selection]
end procedure
```

Data Selection: In this research work, after feature extraction the feature vector is divided into training and testing ratio and sent to classifiers.

Data Classification: Data classification is the process of sorting and categorizing data into various types, forms or any other distinct class. Data classification enables the separation and classification of data according to data set requirements for various objectives. In this research work the proposed algorithm is designed to categorize into three categories.

The initial data set had 61 attributes. The data set is modified by adding a 62nd attribute which is Boolean, named 'Popular' 'Average' and 'Unpopular'. The selected attribute decides the class label of the data set which is based on average and mean of the number of shares which is explained in algorithm below.

Deciding Multi Class of Articles

```
procedure POPULARITY(shares)
for each i in shares do
if i<=median(i)
popularity = 0; // popularity = unpopular
else if i>= average(i)
 popularity = 2; // popularity = popular
 else
popularity = 1; // popularity = average
```

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

end if end if end for end procedure

2. Implementation

In order to evaluate the performance of proposed work, the algorithms are executed and their performances are compared.

2.1 Data Set Description

The dataset is taken from UCI machine learning repository [13]. This dataset is collected from popular news web site known as Mashable.com. It is preprocessed and donated on this UCI repository [2]. Total 61 attributes are extracted from 39,797 news articles and these attributes describe different features of every article. These news articles are collected during 2 years of period, from January 7 2013 to January 7 2015.

2.2 Result Analysis

A news article is popular or unpopular is predicted based on last column of dataset known as 'number of shares' of news article on social media. Threshold value is calculated on 'number of shares' attribute using algorithm discussed above. The entire dataset is split into training and testing set.

In this work, ten prediction algorithms results are analyzed in order to find which algorithm will give us the maximum prediction rate for reduced attributes.

In Table I and figure 2 result analysis is represented and compared with other classifiers.

Table I: Comparative Analysis for Multiclass Classification

Techniques	Accuracy (in %)
Deep Neural Network	76.80
SVM	66.67
KNN	66.30

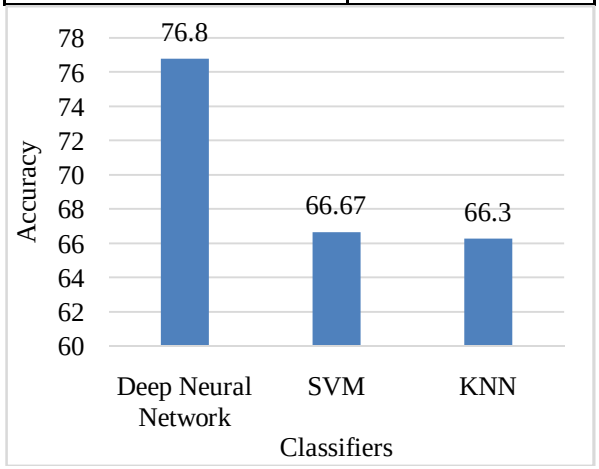

Figure 2: Accuracy Analysis for Multiclass Classification

III. CONCLUSION

News popularity is the maximum growth of attention given for particular news article. Online news popularity depends upon various factors such as number of shares on social media, number of comments by visitors, number of likes etc. So it is necessary to build an automated decision support system to predict the popularity of news as it will help in business intelligence too. In this research work, after applying feature reduction, out of 61 attributes 30 attributes are

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

selected for prediction of popularity of the news. In this work news are predicted into three categories, i.e. popular, average and unpopular. For classification of reduced feature dataset deep neural network, support vector machine and k-nearest neighbor is used. From the experimental results, it is observed that deep neural network outperforms best with respect to support vector machine and k-NN.In this proposed algorithm the work is performed with three classification. So, in future work by analyzing this result, the work is proceeded with more multiclassification.

REFERENCES

[1] Kelwin Fernandes, Pedro Vinagre, Paulo Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News", Springer, EPIA 2015, pp. 535-546, 2015.
[2] He Ren, Quan Yang, "Predicting and Evaluating the Popularity of Online News", Standford University Machine Learning Report.
[3] Bandari Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity." arXiv preprint arXiv:1202.0332, 2012.
[4] IoannisArapakis, B. BarlaCambazoglu, and MouniaLalmas, "On the Feasibility of Predicting News Popularity at Cold Start", Springer, pp. 290-299, 2014.
[5] R. Shreyas, D.M Akshata, B.S Mahanand, B. Shagun, C.M Abhishek, "Predicting Popularity of Online Articles using Random Forest Regression", International Conference on Cognitive Computing and Information Processing, IEEE, 2016.
[6] Joe Maguire, Scott Michelson, "Predicting the Popularity of Social News Posts", Machine Learning Report, Standford University.
[7] Tim Paek, Michael Gamon, Scott Counts, David Maxwell Chickering, Aman Dhesi, "Predicting the Importance of Newsfeed Posts and Social Network Friends", Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10), pp. 1419-1424, 2010.
[8] 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.
[9] UCI Machine Learning Database, https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity, May 2015.