Journal of Theoretical and Applied Information Technology
30th April 2020. Vol.98. No 08

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PREDICTING THE POPULARITY OF ONLINE NEWS USING CLASSIFICATION METHODS WITH FEATURE FILTERING TECHNIQUES

RUBA OBIEDAT
King Abdullah II School of Information Technology, the University of Jordan, Amman, Jordan
E-mail: r.obiedat@ju.edu.jo

ABSTRACT

Due to the expanded use of the internet and the revolution of the information technology field, people are beginning to read news online more and more. For that reason, online news has become the main source of information for the majority of people, and predicting the popularity of online news has become a hot topic as it could help writers present competitive and highly readable news. These predictions can be done by using machine learning techniques. This paper introduces some of the most well-known prediction classification models in data mining like Random Forest, Bayes Net, Logistic Function, C4.5 and Simple Cart which has been applied in order to predict the popularity of the online news. The objective of this paper is to evaluate the performance of these different models on real-world online news data. The experiment results revealed the success of some of these models in predicting the popularity of online news with relatively high accuracy. The performance of the five models is evaluated by some of the most popular metrics such as Accuracy, Root Mean Square Error (RMSE), Kappa Statistic, TP-Rate, FP-Rate, Precision, F-Measure and ROC Area values. Finally, feature filtering techniques are applied in the study to improve the model's performance and identify the most influential features affecting news popularity.

Keywords: *Online News, Classification, Feature Filtering, Popularity, Data Mining*

1. INTRODUCTION

An article is a written work that can be published either in hard copy format (e.g.newspapers) or in soft copy format (e.g. online article) [6]. Due to the ever-growing use of the internet, social networks and technological revolution (e.g. the use of smart-phones) thousands of online news sites were available to online readers [12], so users tend to read online news more than newspapers, which makes online news the main source of information for the major part of the community [15]. Moreover, online news has many features that make them more adopted by news organizations, like the small size, low publication cost, and easy construction [23].

An article is considered to be a popular piece of news if it is among the most read and appealing articles on a particular day of publishing in a certain news outlet [12]. The popularity of an article can be measured based on several factors, such as the number of views, comments, likes, votes, or shares through social networks or by email [8, 27].

The Popularity of news is valuable and useful for many sectors, like business, marketing and online advertising, recommendation systems, and even in political activities, because people prefer reading the most popular articles and sharing it with friends, which is likely to influence public interest and opinions [4]. Many companies in the world spend up to 30% of their budget on online marketing [24], therefor, online news web sites and social media pages compete to attract more readers and try to improve the quality of online articles before their publication in order to attract these companies. Moreover, news sites can employ predictions in order to highlight their popular news [23], and arrange it on their home page accordingly [15], to try to attract the readers by identifying their interest and focusing on the relevant and engaging news that they will find interesting [12]. As a result, online news web sites can allocate their resources better to write stories on the selected topics at the right time [27]. Furthermore, online readers can filter the huge amount of available information quickly and easily, and focus on the most important ones [23, 27]. On the other hand, it can help governments allocate harmful news and stop publishing such news [15]. Many features may influence users' interest in a piece of particular news, such as the topic, timing, length, position on the web page, keywords, or extra media [12].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Because predicting the popularity of online news accurately before publishing would be helpful for online sites workers, this subject is becoming one of the most recent research trend for businesses as it helps in identifying whether a piece of news will capture the readers' attention. Additionally, online advertisement strategies have become increasingly interested in understanding reader behavior and predicting the articles that may gain big user notice [23]. Hence, it is necessary for social networking websites and online news writers to build an automated model that predicts the popularity of news prior to their publication, and such models can be implemented through the use of business intelligence and data mining tools.

Predicting the popularity of online news is a challenging and complex task for many reasons since too many factors affect the popularity of a topic. First, it is hard to measure the quality of the content or its relevance to the readers' interest, besides the inaccessibility of the contexts outside the web, as well as the local and geographical conditions that may influence the population and make the prediction more difficult. Moreover, the relationships between the content and the events in the real world are hard to capture and feed into the prediction engine [27]. Also, the popularity depends totally on user behavior, interest, feelings, and point of view which is very hard to be predicted. Prediction, especially prior to publication, lacks many discriminatory features. In addition, complex social interactions and information cascades make it very hard to predict at the microscopic level. Furthermore, it is hard to analyze the text of the web and semantic data; the structure of the network and the interaction between the different layers of the web presents another challenge [27, 4] and the page structure complexity like the first page location and the second page make the prediction even more difficult. Other unpredictable factors such as social media influencers, can also affect the future of a topic [15].

Data mining models present an appropriate method for this purpose especially with a classification approach since it can label the data into predefined binary classes (e.g. “true”, “false”). Several data mining algorithms can be implemented for predicting the news popularity such as Decision Trees (DT) and Support Vector Machines.

There are two approaches that can be followed by prediction techniques in order to measure news popularity; the first one is after publication, which captures users' attention for certain news items after their publication. This approach expects higher accuracy since it is easier to utilize the features of reader attention after publication, such as click stream and comments information. The other one is prior to publication, which presents a more challenging approach and expects lower results, as it relies on metadata features instead of original news content [27]. This paper aims to develop a model based on machine learning to predict the popularity of online news prior to its publication by using different classification algorithms. The goal is to obtain a model with high predictive power that decision makers can rely upon. Consequently, real dataset has been used from the UCI Machine Learning Repository with 39,644 articles from the Mashables online news website. A binary classification is used to consider whether an article is popular or not based on the number of the article shares across social media sites.

This paper is structured as follows: Section 2 presents the existing works relevant to predicting online news popularity. Section 3 describes the followed methodology based on classification algorithms and discusses evaluation metrics. The results are then analyzed and the performances compared in Section 4. Finally, the conclusion is presented in section 5.

2. RELATED WORKS

Many factors play an important role in the popularity of news on social media platforms and must be taken into consideration, like news content [10] promotion, social influence, connections, and sharing [8]. Several algorithms were implemented for predicting the popularity of online news, but the most popular ones were the Decision Tree and Support vector Machine [12, 19, 17].

Some researchers such as Kelwin et.al. [9] proposed a system called Intelligent Decision Support System (IDSS) that first analyzes an article prior to publishing and predicts its popularity, then it optimizes a set of article features that can be changed by the author to enhance the prediction. The authors collected a large dataset of articles (almost 39,000) articles from the Mashbel news service and the best result was achieved by using Random Forest (RF) with an accuracy of 0.67%. The authors then donated the collected data to the UCI Machine Learning Repository.

A study that was conducted on real dataset from UCI Machine Learning Repository, aimed to find the best model to predict the popularity of online news by using data mining methods, the author used LDA to reduce the dimension as well

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

as three different learning algorithms, such as AdaBoost, LPBoost, and Random Forest. The best result was gained using the Adaptive Boosting model, as it has achieved an accuracy of 69%, and an F-measure of 73% [8]. A study done by Choudhary aiming to maximize the rate of popularity prediction of an article by selecting a minimum number of optimum features. Using dataset from the UCI Machine Learning Repository with 39,644 articles with 60 attributes, and one class label attribute. Genetic algorithm was used to get the best set of attributes that should be considered while constructing an article. The results showed that Naïve Bayes had the best prediction value for 32 attribute set with an accuracy of 93.46%, and Neural Networks had the best prediction value for 18 attribute set with an accuracy of 91.96% [6].

In addition, Alexandru et al. [23] studied the problem of predicting the popularity of news articles according to user comments, and their goal was to accurately rank articles based on their predicted popularity. Authors used data from a renowned French online news platform, their results showed that simple linear regression is the best prediction method as it improved the ranking performance. Moreover, Hensinger et al. [12] tested popularity prediction based on the ‘‘appeal’’ function, were popular articles are considered to be the most appealing on a particular day. The authors used Ranking Support Vector Machines (SVM) along with text features, like keywords to obtain better results for the appeal function. Data was collected from six different outlets over a period of 1 year. SVM gave high computation results. As for Arabic articles on Wikipedia, Hanadi Muqbil, AL-Mutairi, Mohammad Badruddin, Khan used several classifiers, such as; DT, Wjrip, and NB for predicting popularity levels based on stimulant features. They found that the two main external stimulants that are the most important in predicting the popularity of trending Arabic Wikipedia articles are breaking news and annual events [1]. Ioannis Arapakis, B. Barla Cambazoglu, and Mounia Lalmas [2] performed a study on cold-start news popularity prediction, which was conducted on 13,319 news articles acquired from Yahoo News. The popularity was measured based on two metrics: tweet counts and page views. The experimental results of the classifiers used (like NB, J48, and SVM) revealed a bias to learn unpopular articles due to the imbalanced class distribution. Also they indicated that predicting the news popularity at cold-start is a difficult task. This research aims to use other data mining algorithms and make a comparison among them in order to find the best model for predicting the popularity of online news. Furthermore, it highlights the most important features affecting news popularity, which helps editors, writers, and decision makers to enhance popularity and prediction rates.

3. METHODOLOGY

In order to develop the news popularity model, a methodology based on three stages is followed and implemented. The methodology consists of the following steps in order: data acquisition, followed by construction of the classification models and performance evaluation of the models. These stages are described in detail as follows:

Table 1: Part of the available features.

Feature	
WORD	Number of words of the title/article Rate of non-stop/unique words
Links	Number of links; Minimum, average an maximum number
Digital Media	Number of images and videos
Time	Day of week Published on weekend?
Keywords	Number of keyword average of keywords(min./avg ./max.shares)
NLP	Title subjectivity Title sentiment polarity
Target	Number of shares

3.1 Data Acquisition

As mentioned before, the used dataset was obtained from the UCI Machine Learning Repository, which were basically collected and pre-processed by K.Fernandes [9]. The dataset consists of 39,644 articles that have been posted on the Mashable website between the years 2013 to 2015, with a total of 60 attributes and one class label “shares” (as numerical values) to describe the various features of each article. After loading the CSV data file, a new column was created for the popularity (as a nominal value) after the shares column. Using the IF function; the class is popular “Yes”, if the number of shares is greater or equal than 1400, otherwise it is considered as not popular and labeled with “No”. Moreover, URLs and timedelta columns were removed because they were meta-data and cannot be used as characteristics, also we removed the shares column and assigned the popular column to be the class label, which

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

finally led to a total of 59 attributes. One important aspect in the machine learning workflow is to check whether the dataset is imbalanced (unequal distribution of instances between its classes), it was found that the number of instances with the class label “Yes” were equal to 21,154 while instances with class label "No" equal to 21,199. Hence, the used dataset is balanced. Part of the features set is categorized in Table 1.

3.2 Constructing the Classification Models

In this paper, five different prediction algorithms were applied to the same dataset to check their performance in predicting the popularity of online news.

- Random Forest: Random Forest is a classification method that combines many decision trees based on individual sets of examples from the dataset. Each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In the Random Forest algorithm, the results of decision trees are combined to select the most popular class. The Random Forest classifier can be defined by the relation:

$$H(x) = argmax_Y \sum_{i=1}^{k} I(h_i(x) = Y) \quad (1)$$

Where k is the number of decision trees, Y is the class label, and H(x) is the final combined classifier [18]. Random Forest has many advantages that contribute to its popularity, it is easy to implement, classifies correctly, robust, and able to handle a large set of data.

- Bayes Net: Bayesian Network is a classification algorithm that is based on the theorem of Bayes, where the probability of each node is measured, and a Bayesian Network is formed. A Bayesian Network (BN) is a probabilistic graphic model that allows us to manipulate probability distributions effectively [3]. Bayesian Network is a cyclically directed graph that shows a data feature and a set of probability distributions in every node [5]. Probabilistic classification can be performed by Bayes theorem as follows:

$$p(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \quad (2)$$

- Logistic Function: It is a classification algorithm used for predictive problems and it is based on the concept of probability. The logistic function can be defined as the sigmoid function. It is useful to predict the presence or absence of a feature or a result based on the values of a series of predictor variables. It is similar to a linear regression model but suitable for models that have a dichotomous dependency variable [14]. The sigmoid method was used to model expected probability values, as the method maps every true value to a different value from 0 to 1.

- SimpleCart: It is a prediction algorithm developed by Leo Breiman in the early 80s. It relies on historical data in order to build a decision tree, and it works with either categorical or numerical attributes that differentiate it from other decision tree algorithms [21]. CART can handle outliers as it isolates them in individual nodes during the splitting process. The CART algorithm adapts the following steps: it first constructs the maximum tree size, then finds the right tree size, and finally makes the classification using a constructed tree [25].

- C4.5: It was developed by Ross Quinlan and is used to generate a decision tree [11]. It is an extension of the ID3 algorithm that aims to overcome most of the ID3 weaknesses such as dealing with noise and missing data. C4.5 builds a decision tree based on the information gain concept. The attribute with the highest information gain is placed as the splitting node. Furthermore, C4.5 uses Gain Ratio for the attribute selection criteria. This method contains two concepts, Gain and Split Info [22].

3.3 Data Acquisition

In order to evaluate the performance of the proposed algorithms on the same online news

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dataset, Weka 3.9.3 was used for the evaluation process. Weka is a popular open-source suite of machine learning software written in Java and developed at the University of Waikato, New Zealand. Weka has a simple interface to operate with, and it can be implemented on any platform. Weka supports various tasks of data mining like data pre-processing, clustering, classification, regression, visualization, and feature selection. In our experiment we implemented Weka on windows 10 with 8 GB RAM. Moreover, for dividing the dataset into a training and testing set, 10-fold cross-validation was used. 10-fold cross validation is a common method in machine learning that is used to evaluate machine learning models; it randomly divides the dataset into 10 groups (folds) of approximately equal size. The first fold is treated as a validation set, and remaining the 9 folds are treated as a training set. This process is repeated 10 times, and then the model's accuracy is calculated as the average of the obtained accuracy in each round [26]. The measurement in this paper of the experimental result is evaluated by using the Confusion Matrix, which is a matrix that describes the performance of a classification model.

Each row of the matrix represents the number of instances in a predicted class while each column reflects the instances in an actual class (or vice versa). Confusion Matrix for binary classifier consists of four categories (as number not rate):

- True Positive (TP): indicates the positive prediction and it is true.
- True Negative (TN): indicates the negative prediction and it is true.
- False Positive (FP): indicates the positive prediction and it is false.
- False Negative (FN): indicates the negative prediction and it is false.

From the Confusion Matrix the Recall, Precision, ROC (Receiver Operating Characteristics), and Accuracy can be obtained easily, and are used to measure the performance of the machine learning classification models [26, 7].

Recall as it is also known as True Positive Rate (TPR) measures the fraction of positive instances that are correctly labeled [7]. It is given by the relation:

$$recall = \frac{(TP)}{(TP + FN)} \quad (3)$$

Precision measures the fraction of instances classified as positive that are truly positive [7]. It is given by the relation:

$$precision = \frac{(TP)}{(TP + FP)} \quad (4)$$

Accuracy measures the rate of the correctly classified instances. It is given by the relation:

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

ROC Area is also one of the most popular and important evaluation metrics for checking the performance of any classification model, and it indicates how much the model can differentiate between classes. The ROC curve is plotted with True Positive Rate (Y-Axis), against False Positive Rate (X-axis). An optimal model is with the value of ROC = 1. Moreover, Kappa Statistics and Root Mean Square Error (RMSE) were used to evaluate the performance of the five different models. Cohen's Kappa is a scalar meter of accuracy that was first used as a measure of agreement between observers of psychological behavior. Later, it was found that Cohen's Kappa can also be used as a meter for classifiers' accuracy [20]. Cohen's Kappa is defined by the relation:

$$k = \frac{\mathrm{P}_a - \mathrm{P}_e}{1 - \mathrm{P}_e} \quad (6)$$

Where P_a is the actual agreement probability, and P_e is the expected outcome probability, which is due to chance. Cohen's Kappa ranges from (0-1).

RMSE is frequently used in climatology, forecasting, and regression analyses to verify experimental results. It is a measure of the differences between the predicted values by a model and the true values. RMSE is the standard deviation of the prediction errors that measure how far they are from the regression line data points.

Journal of Theoretical and Applied Information Technology
30th April 2020. Vol.98. No 08

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. RESULTS AND DISCUSSION

After loading the dataset to Weka and applying the five different classification algorithms (Random Forest, Bayes Net, Logistic Function, Simple Cart, and C4.5), we obtained the results of the Confusion Matrix of each classifier illustrated in the tables below:

Table 2: Random Forest Accuracy Results.

		Predicted	
		Yes	No
Actual	Yes	True Positive 15232	False Negative 5922
	No	False Positive 7245	True Negative 11245

Table 3: Bayes Net Accuracy Results.

		Predicted	
		Yes	No
Actual	Yes	True Positive 14671	False Negative 6483
	No	False Positive 7383	True Negative 11107

Table 4: Logistic Function Accuracy Results.

		Predicted	
		Yes	No
Actual	Yes	True Positive 14768	False Negative 6386
	No	False Positive 7359	True Negative 11131

Table 5: Simple Cart Accuracy Results.

		Predicted	
		Yes	No
Actual	Yes	True Positive 14980	False Negative 6174
	No	False Positive 7765	True Negative 10725

Table 6: C4.5 Accuracy Results.

		Predicted	
		Yes	No
Actual	Yes	True Positive 13058	False Negative 8096
	No	False Positive 8129	True Negative 10361

Table 7: Performance Comparison of the five classifiers.

Classifier	Accuracy	Kappa Statistics	RMSE
Random Forest	**66.7869%**	0.3297	0.4586
Bayes Net	65.0237%	0.2951	0.5141
Logistic Function	65.3239%	0.301	0.4659
SimpleCart	64.8396%	0.2898	0.4728
C4.5	59.0733%	0.1777	0.6133

It is noted that the number of the correctly classified instances (TP+TN) of the Random Forest are 26,477, while the number of the correctly classified instances of Bayes Net are 25,778, 25,899 for Logistic Function, 25,705 and 23,419 for Simple Cart and C4.5 respectively from the total 39,644. Firstly, a comparison was conducted between these classifiers based on the three different performance measures, as shown in Table 7.

An important component in developing classifiers to predict the popularity of online news is the ability to determine the accuracy of these classifiers by measuring the ratio of the total number of correctly-classified instances, to the total number of instances. From Table 7, each of the classification algorithms showed some difference in the average Accuracy. However, it was found that the Accuracy of Random Forest, Bayes Net, Logistic Function, Simple Cart and C4.5 are 66.7869%, 65.0237%, 65.2389%, 64.8396%, and 59.0733% respectively. The accuracy results were given directly from Weka without the needs of any manual calculations. Hence, it is clear that the performance of Random Forest is better than other classifiers here. Moreover, based on the reported results in Table 7, Random Forest comes out first with a Kappa statistic value of 0.3297 and last of RMSE value of 0.4586, followed by Logistic Function having a Kappa statistic value of 0.301 and a RMSE value of 0.4659, C4.5 stands last with the lowest Kappa statistic value 0.1777 and the highest RMSE value 0.6133. Therefore, with respect to the classification Accuracy as a performance measure Random Forest is the best among all other models followed by Logistic Function, Bayes Net, Simple Cart, and finally C4.5. Figure 1 shows a performance comparison between the five classifier models.

Moreover, these models were compared based on the Recall (TPR), False Positive Rate (FPR), Precision, F-Measure, and ROC area values. All these values were obtained from the Confusion Matrix.

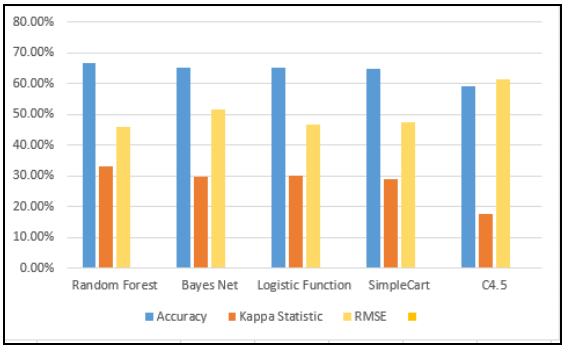

Figure 1: Performance comparison among the different classifiers

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 8: Detailed performance by each of the five classifiers.

Classifier	TP-Rate/Recall	FP-Rate	Precision	F-Measure	ROC
Random Forest	0.668	0.34	0.667	0.667	0.728
Bayes Net	0.65	0.356	0.649	0.65	0.703
Logistic Function	0.653	0.353	0.653	0.653	0.707
Simple Cart	0.648	0.36	0.647	0.647	0.684
C4.5	0.591	0.413	0.591	0.591	0.582

From Table 8, it is noted that the weighted average values of TPR, FPR, Precision, F-Measure and ROC for some methods are quite high. The Random Forest classifier obtained values of 0.668, 0.340, 0.667, 0.667, and 0.728, respectively. Whereas for the Bayes Net classifier the values obtained were 0.650, 0.356, 0.649, 0.650, and 0.703, respectively. For the Logistic Function the values were 0.653, 0.353, 0.653, 0.653, and 0.707, respectively. And for the Simple Cart the values were 0.648, 0.360, 0.647, and 0.684, respectively. In the meanwhile, the prediction ability of C4.5 was the worst among all the classifiers with values of 0.591, 0.413, 0.591, 0.591, and 0.582, respectively. Therefore, the C4.5 algorithm was rejected and finally the Random Forest model was adopted, which has the highest weighted average values for TPR, Precision, F-Measure and ROC area, and the lowest weighted average value for FPR.

4.1 Results Based on Feature Selection Methods

Weka tool has an attribute selection option. Attribute selection is the process of removing irrelevant attributes of the data mining task. It also aims to find the main attributes that really affect the classification results. Weka was used to reduce the number of attributes in order to increase the Accuracy, and since Random Forest is the best model with regards to performance among all other proposed models, it was chosen to apply the feature filtering techniques using the top 50, 40, 30, and 20 features which was selected by the feature selection algorithms respectively. Five different selection attribute methods were applied to the dataset, which are:

- InfoGainAttributeEval: which measures the information gain of the attribute as the main indicator for the relevance of an attribute with respect to the class label.
- ChiSquaredAttributeEval: which computes the Chi-Squared statistic of the attribute as the main indicator for the relevance of an attribute with respect to the class label.
- CorrelationAttributeEval: which measures the Pearson's Correlation between it and the class label as the main indicator for the relevance of an attribute with respect to the class label.
- Gain Ratio: which evaluates the value of an attribute by measuring the gain ratio with respect to the class.
- OneRAttributeEval: One-R is a simple algorithm proposed by Holte [13]. It defines one rule in the training data for each attribute, and then selects the rule with the smallest error. It can handle the missing values by treating "missing" as a legitimate value. This is one of the most primitive schemes, as it produces simple rules based on one feature only. Although it is a minimal form of the classifier, it can be useful for determining a baseline performance as a benchmark for other learning schemes [16].

Table 9 shows the performance metrics of the Random Forest with 10-fold cross-validation and the top number of features which was selected by the feature selection methods in Weka. According to the table, the highest Accuracy value was obtained by the Information Gain, and was 66.9029% for the top 50 selected attributes, which means a small improvement in the Accuracy compared to the total 59 attributes (0.116%). The values of Accuracy, Precision, Recall, and F-Measure were not improved beyond 67% in the existing research work [9]. Moreover, Table 9 shows that the performance of the Random Forest according to Accuracy, Recall, Precision, F-measure and ROC values decreased when the number of selected attributes was reduced.

Additionally, in this part of the experiment the most important 10 attributes were extracted. This step is very important to give more insight to editors and writers, and to assist them in making efficient decisions, and to reveal the most important features that optimize the popularity of the article. The keyword feature (kw-avg-avg,kw-min-avg, and kw-max-avg) plays an important role for increasing the average of shares, followed by Latent Dirichlet Allocation LDA_00, LDA_01, and LDA_02, self-referenced articles in Mashable, and finally the world channel category whether it is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

a lifestyle, business, entertainment, social media, technology or world is considered an important feature to take into consideration.

Table 9: Evaluation of Feature Selection Methods

Attribute selection methods	Accuracy	Recall	Precesion	F-Measure	ROC-Area
Top 50					
Correlation	66.82%	0.668	0.667	0.667	0.728
Information Gain	66.90%	0.669	0.668	0.668	0.729
ChiSquare	66.77%	0.668	0.667	0.667	0.728
One-R	66.64%	0.666	0.666	0.665	0.727
Gain Ratio	66.81%	0.668	0.667	0.667	0.728
Top 40					
Correlation	66.56%	0.666	0.665	0.665	0.726
Information Gain	66.70%	0.667	0.666	0.666	0.728
ChiSquare	66.73%	0.667	0.667	0.666	0.728
One-R	66.24%	0.662	0.662	0.661	0.723
Gain Ratio	66.65%	0.667	0.666	0.666	0.728
TOP 30					
Correlation	66.10%	0.661	0.660	0.660	0.720
Information Gain	66.09%	0.661	0.660	0.660	0.718
ChiSquare	66.13%	0.661	0.660	0.660	0.720
One-R	65.41%	0.654	0.653	0.653	0.709
Gain Ratio	66.61%	0.666	0.665	0.665	0.724
Top 20					
Correlation	65.29%	0.653	0.652	0.652	0.705
Information Gain	65.65%	0.656	0.656	0.656	0.712
ChiSquare	65.71%	0.657	0.656	0.656	0.714
One-R	64.41%	0.644	0.643	0.643	0.696
Gain Ratio	66.33%	0.663	0.663	0.662	0.718

4.2 Comparison With the Existing Work

After analyzing and evaluating the proposed five algorithms, a comparison with the existing research work done by [9] was implemented. The results obtained by previous work using the same online news data set of the Mashable website are illustrated in Table 9. It is observed that Random Forest shows better performance with respect to Accuracy, Precision, Recall, F1 and AUC values (0.67, 0.67, 0.71, 0.69, and 0.73, respectively) among all other models. It can also be noted that we gained very competitive results for RF using the top 50 features based on the information gain feature selection techniques as shown in Table 10 with (0.669, 0.668, 0.669, 0.668, and 0.729, respectively).

In more detail, when comparing Bayes classifiers family, Naive Bayes from the previous work obtained an Accuracy value of 0.62 while the Bayes Net in the presented work obtained accuracy 0.65 which is relatively considered much better for classifying the instances correctly. Moreover, the presented algorithms (i.e. Simple Cart and Logistic Function) obtained better results with respect to Accuracy (0.648 and 0.653, respectively) than the K-Nearest Neighbors (0.62) in the previous work. Therefore, it is found from the obtained results of both works that Random Forest is the best model that can be used for predicting the popularity of online news.

Table 10: Comparison of models performance [9].

Model	Accuracy	Precision	Recall	F1	AUC
Random Forest (RF)	0.67	0.67	0.71	0.69	0.73
Adaptive Boosting (AdaBoost)	0.66	0.68	0.67	0.67	0.72
Support Vector Machine (SVM)	0.66	0.67	0.68	0.68	0.71
K-Nearest Neighbors (KNN)	0.62	0.66	0.55	0.60	0.67
Naıve Bayes (NB)	0.62	0.68	0.49	0.57	0.65
IG+RF (proposed)	0.669	0.668	0.669	0.668	0.729

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION

News popularity prediction is becoming an important topic nowadays because of the huge expansion of online news, smart-phones, and web2. Online news has become the main source of information for the majority of people, as it provides a piece of valuable information for many sectors such as business, marketing, politics, and social media. This paper focused on using procedures to evaluate and compare the performance of five classification models which are Random Forest, Bayes Net, Logistic Function, Simple Cart and C4.5 on the online news dataset. The classifications basically aim to predict whether an article is to be popular or unpopular, based on the number of shares, where the threshold is set to 1,400. The performance of these five classification models has been evaluated using several of the most common and popular evaluation metrics in the data mining field. Results show that Random Forest was the best classifier compared to other presented models, since it has the highest Accuracy value of 66.7869% and 0.3297 Kappa Statistics, and the lowest RMSE value of 0.4586. Hence, the Random Forest could be very much helpful for online news popularity prediction. Different Feature selection methods were applied to RF since it presents the best classifier such as InfoGainAttributeEval, ChiSquaredAttributeEval, CorrelationAttributeEval, Gain Ratio, and OneRAttributeEval. It was checked against the top 50, 40, 30, and 20 features. Results show a small improvement in Accuracy using the top 50 features. Another interesting finding concerning the top 10 features which can give decision makers an essential insight for the main features affecting the popularity to focus on.

REFRENCES:

[1] 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." In 2015 *International Conference on Cloud Computing* (ICCC), pp. 1-6. IEEE, 2015.

[2] Arapakis, Ioannis, B. Barla Cambazoglu, and Mounia Lalmas. "On the feasibility of predicting news popularity at cold start." In International *Conference on Social Informatics*, pp. 290-299. Springer, Cham, 2014.

[3] Arias, Jacinto, Jose A. Gamez, and Jose M. Puerta. "Learning distributed discrete Bayesian network classifiers under MapReduce with Apache spark." *Knowledge-Based Systems* 117 (2017): 16-26.

[4] Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity." In Sixth International AAAI *Conference on Weblogs and Social Media*. 2012.

[5] Bozkurt, Sinem, Gulin Elibol, Serkan Gunal, and Ugur Yayan. "A comparative study on machine learning algorithms for indoor positioning." *In 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, pp. 1-8. IEEE, 2015.

[6] Choudhary, Swati, Angkirat Singh Sandhu, and Tribikram Pradhan. "Genetic algorithm based correlation enhanced prediction of online news popularity." *In Computational Intelligence in Data Mining*, pp. 133-144. Springer, Singapore, 2017.

[7] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." *In Proceedings of the 23rd international conference on Machine learning*, pp. 233-240. 2006.

[8] Deshpande, Dhanashree. "Prediction & evaluation of online news popularity using machine intelligence." *In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1-6. IEEE, 2017.

[9] Fernandes, Kelwin, Pedro Vinagre, and Paulo Cortez. "A proactive intelligent decision support system for predicting the popularity of online news." *In Portuguese Conference on Artificial Intelligence*, pp. 535-546. Springer, Cham, 2015.

[10] Figueiredo, Flavio, Jussara M. Almeida, Fabrício Benevenuto, and Krishna P. Gummadi. "Does content determine information popularity in social media? a case study of youtube videos' content and their popularity." *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 979-982. 2014.

[11] Gupta, Bhumika, Aditya Rawat, Akshay Jain, Arpit Arora, and Naresh Dhami. "Analysis of various decision tree algorithms for classification in data mining." *Int. J. Comput. Appl* 163, no. 8 (2017): 15-19.

[12] Hensinger, Elena, Ilias Flaounas, and Nello Cristianini. "Modelling and predicting news popularity." *Pattern Analysis and Applications* 16, no. 4 (2013): 623-635.

[13] Holte, Robert C. "Very simple classification rules perform well on most commonly used

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

datasets." *Machine learning* 11, no. 1 (1993): 63-90.

[14] Kurt, Imran, Mevlut Ture, and A. Turhan Kurum. "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease." *Expert systems with applications* 34, no. 1 (2008): 366-374.

[15] Liu, Caiyun, Wenjie Wang, Yuqing Zhang, Ying Dong, Fannv He, and Chensi Wu. "Predicting the popularity of online news based on multivariate analysis." *In 2017 IEEE International Conference on Computer and Information Technology (CIT)*, pp. 9-15. IEEE, 2017.

[16] Novaković, Jasmina. "Toward optimal feature selection using ranking methods and classification algorithms." *Yugoslav Journal of Operations Research* 21, no. 1 (2016).

[17] Paek, Tim, Michael Gamon, Scott Counts, David Maxwell Chickering, and Aman Dhesi. "Predicting the importance of newsfeed posts and social network friends." *In Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.

[18] Pu, Xiaorong, Ke Fan, Xiong Chen, Luping Ji, and Zhihu Zhou. "Facial expression recognition from image sequences using twofold random forest classifier." *Neurocomputing* 168 (2015): 1173-1180.

[19] Ren, He, and Quan Yang. "Predicting and Evaluating the Popularity of Online News." *Standford University Machine Learning Report* (2015).

[20] Rosenberg, Andrew, and Ed Binkowski. "Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points." *In Proceedings of HLT-NAACL* 2004: short papers, pp. 77-80. 2004.

[21] Sallis, Philip J., William Claster, and Sergio Hernández. "A machine-learning algorithm for wind gust prediction." Computers & geosciences 37, no. 9 (2011): 1337-1344.

[22] Sathyadevan, Shiju, and Remya R. Nair. "Comparative analysis of decision tree algorithms: ID3, C4. 5 and random forest." *In Computational intelligence in data mining-volume 1*, pp. 549-562. Springer, New Delhi, 2015.

[23] Tatar, Alexandru, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. "Ranking news articles based on popularity prediction." *In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 106-110. IEEE, 2012.

[24] Tatar, Alexandru, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. "A survey on predicting the popularity of web content." *Journal of Internet Services and Applications* 5, no. 1 (2014): 8.

[25] Timofeev, Roman. "Classification and regression trees (CART) theory and applications." *Humboldt University, Berlin* (2004): 1-40.

[26] Turban, Efraim, Ramesh Sharda, and Dursun Delen. *Business intelligence and analytics: systems for decision support*. Pearson Higher Ed, 2014.

[27] Uddin, Md Taufeeq, Muhammed Jamshed Alam Patwary, Tanveer Ahsan, and Mohammed Shamsul Alam. "Predicting the popularity of online news from content metadata." *In 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 1-5. IEEE, 2016.