International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

A Comparative Analysis of Machine Learning Techniques for Online News Popularity Prediction

Shweta Jagtap[1], Saanya Lakhani[2], Akash Devadiga[3]

[1-3]*Department of Information Technology, Atharva College of Engineering*

Abstract - *With the expansion of the Internet, more and more people enjoy reading and sharing online news articles. Most people nowadays have switched to online mode for getting their daily news. As news articles are shared over the internet through social media platforms and brought to the attention of many people, they are likely to gain popularity. Predicting news popularity prior to publication has proved to be useful for online news publishers. Various works have been done for predicting online news popularity using different machine learning methods. This research work is based on comparative analysis of such machine learning techniques and choosing the most suitable technique for the problem of predicting the popularity of online news articles.*

Keywords: Machine Learning, Classification, Data Pre-processing, Online news popularity, Ensemble learning.

1. INTRODUCTION

In this era of internet and digitization, a huge amount of information is being made available online, including news. Since the introduction of web 2.0, the practice of producing and sharing online content has gained an increasing amount of attention. There has also been a growing interest in online news, which allows an easy and fast spread of information around the globe. There are various news publishers like Mashable and TechCrunch, publishing their news articles on the internet to engage with a wider range of readers. And if their news article appeals to a larger population, it is shared by many on social media. For the online news websites, their content becoming popular brings in various opportunities for them in terms of business.

As the amount of content being published online has increased, so has the competition among online news and content providers, aiming for an increase in readership. For such news publishers a way of predicting popularity of their news articles can prove to be extremely beneficial.

For building a news popularity prediction system, machine learning techniques prove to be extremely useful. It is necessary to understand the different factors affecting the popularity as well as picking the most suitable algorithms for this particular problem and doing necessary testing and evaluation. In this work, a comparative analysis of various machine learning techniques has been done for prediction of popularity of online news as well as for selecting the
best features which contribute the most towards the prediction.

The remaining paper is organized as follows. Section 2 contains the review of literature. Section 3 includes the comparative analysis of machine learning techniques. Section 4 contains future scope and Section 5 is the conclusion of the paper.

2. REVIEW OF LITERATURE

In machine learning the performance of a model depends on various factors including the data and the features used for training the model. This was highlighted in work [1], where the authors focused on selecting the right features prior to training their model. They performed Univariate selection, Feature Importance and Recursive Feature Elimination (RFE). Due to the large number of features available, Principal Component Analysis (PCA) was also implemented with the aim of reducing the number of features. But the results obtained were not useful enough and it was concluded that the data is not linearly separable. Some of the machine learning algorithms implemented included Naive Bayes, which performed poorly, the reason for this, according to the authors, was attributed to high correlation among the data set features. Logistic Regression, gave a higher performance and it was found that the problem could be solved better by non-linear means. Support Vector Machine was found to be more time and space consuming; Neural Network was able to accurately model the non-linearity. Ensemble models like Random Forest, Gradient Boosting, Bagging and AdaBoost ensemble models gave a significantly better performance. The Gradient Boosting model was the best performing one with 79.7% accuracy.

In work [2], the authors focused mainly on the underlying sentiments in headlines of the news articles. In this paper, sentiment analysis on the data was performed using VADER (Valence Aware Dictionary and Sentiment Reasoner). K-means clustering was implemented to obtain 2 separate clusters of popular and unpopular news. To overcome the skewness of class distribution, cluster-based oversampling was performed. For prediction, Logistic Regression was implemented which was used to set a bar for more complex models like the Artificial Neural

Network (ANN) and XGBoost. With some parameter tuning and cross validation, XGBoost outperformed the Artificial Neural Network achieving maximum accuracy.

In paper [3], the authors proposed a two-stage method for online news popularity prediction. For data pre-processing, word segmentation and stop-word elimination were performed, the natural language was represented as computer language using the Vector Space Model (VSM) with TF-IDF (Term Frequency Inverse Document Frequency) weights, category label and popularity label. The features were extracted using Bag of Words (BoW). The first stage involved selecting the global features which are related to news column information from all news content. Feature selection using ANOVA (Analysis of Variance) for each feature (unique word) was done and the news was classified into different columns. In stage two, the authors created a VSM for each column of news. Local features related to news popularity were selected. The predictive models like SVM, K nearest neighbours (KNN), Naive Bayes, Random Forest, Gradient Boosting and AdaBoost were implemented. Naive Bayes gave the best performance for all the news categories and especially for social and financial news which were, as the authors of this paper claimed, more predictable than any other news category.

In this paper [4] the authors worked with many data mining algorithms. For classification, Neural Networks, Random Forest, Support Vector Machines (SVM), Naïve Bayes, K-Nearest Neighbours, Linear Regression, Logistic Regression, AdaBoost, Bagging were implemented. Precision, Recall, and F-measure were used for evaluation and their results were compared to find the best one. Random Forest and Neural Network were found to be better performing with an accuracy of nearly 65% with optimal parameters.

In paper [5], various methodologies were studied and analysed for predicting popularity of online news as well as some improvements were suggested for the unpopular news articles. In light of the content of the popularity ranking model, back-propagation neural networks (BPNN) were taken to predict popularity using artificial neural networks. The mock-up solutions compare many forecasting methods based on some factors achieved in previous work. This provides a successful prediction model according to real situations, with an accuracy level of 95%.

In work [6], a Deep Fusion of Temporal process and Content features (DFTC) method to handle them were presented. For modelling the temporal popularity process, the authors considered recurrent neural networks and convolutional neural networks. For multi-modal content features, hierarchical attention network and the embedding techniques were used. The proposed model of Temporal Process Modelling and Feature based Modelling, was shown to surpass state of-the-art approach on prediction of popularity.

In paper [7], the authors took the day on which the news are published, the category in which the news belonged and the features of article into consideration. Also, the authors formulated the problem as a binary classification problem and implemented three classification learning algorithms including Logistic Regression, Random Forest and Adaboost. Logistic Regression was used to predict the probability of a categorical dependent variable. Random forest was used as it has lower risk of over fitting. The following metrics were used for evaluation of algorithms- Accuracy, Precision and Recall and AUC. Finally, it was concluded that out of the three classification algorithms Adaboost was the best algorithm for predicting online news popularity.

In Paper [8], different methodologies were analysed which predict the popularity of online news articles.
These methodologies were compared, their parameters were considered and improvements were suggested. The metrics which were computed were accuracy, precision, recall, F1 and AUC. Data was collected from UCI machine learning repository and feature selection was performed on it. LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis) were used for dimensionality reduction. To predict the popularity of online news various classification methods were used. It was found out that Random Forest was the best model for prediction as it showed the best result among other various classification methods. It was also suggested that to improve the value of accuracy, neural networks will be studied and compared.

The work presented in paper [9] intends to find the best prediction model to predict the popularity of online news by using Machine Learning methods. Three different learning algorithms such as Adaboost, LPboost, and Random Forest were implemented and their performances were compared. The prediction performances of all these three methodologies were studied and it was found out that Adaptive Boosting turns out to be the best model for prediction as it achieved an accuracy of 69% and F-Measure of 73%. In this paper, it was suggested that the evaluation parameters such as F-measure and accuracy for popularity prediction can be further improved by natural language processing tools and techniques to understand the semantics of text.

3. COMPARATIVE ANALYSIS

3.1 Data Collection and Data Pre-processing techniques

Data collection is the process of acquiring data for various purposes, like solving a domain specific problem. The data can be gathered from many sources which provide previously collected datasets donated by various researchers or can be obtained by scraping data off the internet, by conducting surveys or using experimental results. The method of data collection can depend on the need or the problem at hand or preference of the researcher. Collecting data is the first and most important step in machine learning.

The raw data obtained from data collection might not always be in the required format. To prepare the data to fit to the machine learning algorithms in further steps, the data needs to be cleaned by removing any outliers or missing values, transformed and encoded in case categorical variables or textual data is present. Feature engineering or feature selection techniques can also be used to deal with unwanted or redundant features. This may also help in improving the machine learning model's performance.

One-Hot Encoding

Machine learning algorithms do not understand textual data, they only accept numeric values. So, the categorical variables in a dataset need to be converted to numerical format. One hot encoding is one such technique which is used to convert the various categories into separate columns, replacing the values with 1 or 0 only.

Advantages: Unlike some other encoding techniques, one hot encoding does not simply map numbers (1,2,3…) to categories, therefore, no ordinal relationship will be assumed by the model during training.

Disadvantages: If one-hot encoding results in high cardinality, then it might lead to the problem of the curse of dimensionality.

Principal Component Analysis

Principal Component Analysis is a dimensionality reduction technique which is used to generate new features (principal components) containing the most information from existing features in large datasets. PCA involves – standardization of data, generating covariance matrix of the dataset variables, computing eigenvectors (direction of axes with most information) and eigenvalues (amount of variance) form the covariance matrix for identifying the principal components. The principal components are constructed in such a way so as to get the highest possible variance (most information).

Advantages: The principal components computed will be linearly independent of one another as well as contain the most information obtained from all the original variables.

Disadvantages: One disadvantage of PCA is that the principal components obtained are not clearly interpretable.

Recursive Feature Elimination

Recursive Feature Elimination is a wrapper-based feature selection technique, which means that it will use different machine learning algorithms at its core and recursively use different subsets of dataset features to fit the model. It uses filter-based selection internally. It recursively keeps on removing the weak features until the specified number of features is reached. In RFE, each feature is ranked by the model's feature_importance_ or coef_ attribute. RFE is used to get a combination of attributes which contribute the most towards prediction of the target variable.

Advantages: Recursive Feature Elimination is easily configurable. RFE is also really effective at selecting the most relevant features for prediction.

Disadvantages: Since it recursively eliminates less important or redundant features, it can be computationally expensive.

Pearson Correlation

Pearson's correlation is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score.

Advantages: Advantage of Pearson correlation is that it is simple to apply.

Disadvantages: Pearson correlation is not specific and accurate.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is used to perform 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 calculation.

Advantages: Guarantees maximum separation, maximizes the relationship between the class variance and the class variance in the given record.

Disadvantage: Newer algorithms have much better prediction as compared to LDA as it is an older algorithm.

3.2 Predictive Modelling

Predictive modelling stage involves using machine learning algorithms for predicting outcomes and hence making better decisions. There are various algorithms available to choose from. The selection of a machine learning algorithm depends on the use case, type of problem to be solved and some other factors like training time, amount of memory required, etc. Selecting the most suitable algorithm for solving any problem, taking into consideration all the requirements and limitations is extremely important.

Gradient Boosting

Gradient Boosting is an ensemble machine learning algorithm. It uses the boosting technique for prediction which is based on the idea of combining weak learners with strong learners, these learners are typically decision trees. Gradient descent is used to minimize the loss (error) function. It is the partial derivative of loss function. In gradient boosting the subsequent models learn from the mistakes of the previous models. Finally, all the predictors are combined by assigning some weight to each in order to get the overall prediction.

Advantages: An advantage of Gradient Boosting is that it provides a better accuracy. Also, gradient boosting is very flexible and provides several hyperparameter tuning options.

Disadvantages: A disadvantage of Gradient Boosting is that it is computationally expensive compared to some other algorithms.

XGBoost

Extreme Gradient Boosting or XGBoost is an ensemble technique. It is a specific implementation of gradient boosting algorithms. It optimizes gradient boosting using system optimizations like parallelization of tree construction for utilizing all CPU cores, cache optimization of data structures and algorithm, tree pruning, etc. It also makes certain improvements in the algorithm like regularization for penalizing complex models to reduce overfitting, cross-validation, automatic handling of missing values in the dataset.

Advantages: XGBoost model takes less amount of time for training and still gives a high performance. It also has the capability of handling missing data and has in-built cross validation.

Disadvantages: A limitation of XGBoost is that it's sensitive to overfitting. Also, it is harder to perform hyperparameter tuning due to the presence of a huge number of hyperparameters.

Naïve Bayes

Naïve Bayes is a classification technique based on Bayes theorem. It assumes that the presence of a particular feature in a class is unrelated to presence of any other feature. There are 3 types of naïve bayes classifiers – Bernoulli naïve Bayes which assumes that all the features are binary, Multinomial naïve Bayes which is used when there are discrete values in data and Gaussian naïve Bayes which assumes normal distribution of data.

Advantages: Naïve Bayes algorithm is fast and easy to implement. It also performs well for multiclass classification problems. It is scalable with large datasets.

Disadvantages: Naïve Bayes assumes that the predictors are independent which is usually not the case in real life scenarios.

Random Forest

Random Forests (RF) is an ensemble classification algorithm which uses trees as base classifiers. The philosophy behind classifier ensembles is based upon the premise that a set of classifiers do perform better classifications than an individual classifier.

Advantages: Random forest can solve both types of problems that is classification and regression and does a decent estimation at both fronts.

Disadvantages: It does a good job at classification but not as for regression problems as it does not give precise continuous nature prediction. It may over fit data sets that are particularly noisy.

AdaBoost

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique which is one of the Ensemble Methods in Machine Learning. In Adaptive Boosting the weights are re-assigned to each instance, with higher weights to incorrectly classified instances. Boosting can be used to reduce bias as well as the variance for supervised learning. Adaboost works on the principle where learners are grown sequentially. Adaboost also works on the same principle as boosting but there is some difference in working of both.

Advantages: The most important advantages of AdaBoost are low generalization error, easy to implement, works with a wide range of classifiers, no parameters to adjust.

Disadvantage: Sensitive to outliers and noisy data. Weak classifiers being too weak can lead to overfitting and low margins.

Neural Network

A neural network is a collection of interconnected perceptron arranged in different layers (input layer, hidden layers and output layer). Each neuron in a neural network has inputs, weight and activation function. The structure of a neural network tries to replicate the human brain and biological structure of neurons. Neural Networks can be used to build predictive models using large datasets. It is used to recognize the underlying relationship among data to generate the best possible solution for a problem.

Advantages: Neural Networks can learn and model complex non-linear relationships, it also has fault tolerance and the ability to generalize.

Disadvantages: Neural Networks require processing and parallel processing functionality for training, the functioning of neural network to get the desired output cannot be explained.

Table - 1: Comparative Analysis of Machine Learning Models

Paper	Data Source	Model	Results
News Popularity Prediction Using Ensemble Methods of Classification [1]	UCI Machine Learning Repository Online News Popularity Dataset	Gradient Boosting	Accuracy- 79.9 %
Comparative Analysis of Statistical Classifiers for Predicting News Popularity on Social Web [2]	Multi-source dataset using social media sources (Facebook, Google+, LinkedIn)	XGBoost	Accuracy- Facebook: 84.31 % Google+: 75.43 % LinkedIn: 99.89 %
A Two Stage Prediction Method for News Popularity Using Content Features [3]	Website		
Online News Popularity Prediction [4]	Dataset collected from UCI machine learning repository.	Random Forest	Accuracy - 65.8%
Popularity Prediction of online news based on radial basis function neural network with factor methodology [5]	UCI machine learning repository online news popularity dataset	Neural Network	Accuracy – 95%
Prediction & Evaluation of online news popularity using machine intelligence [9]	Dataset collected from UCI machine learning repository.	AdaBoost	Accuracy – 73%

3.3 Model Evaluation

To find the most suitable algorithm for a particular problem we use certain evaluation metrics to check the effectiveness of the trained machine learning models. For the evaluation metrics considered, following terminologies are used –

- **TP (True Positive)-** This refers to the positive tuples that were correctly labeled by the classifier.
- **TN (True Negative)-** This refers to the negative tuples that were correctly labeled by the classifier.
- **FP (False Positive)-** This refers to the negative tuples that were incorrectly labeled as positive by the classifier.
- **FN (False Negative)-** This refers to the positve tuples that were incorrectly labeled as negative by the classifier.

Table – 2: Model Performance Evaluation Metrics

Evaluation metric	Description	Formula
Accuracy	Accuracy is the number of correct predictions made by the classifier.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Precision gives us the proportion of data points which were actually correctly predicted by the classifier.	$\frac{TP}{TP + FP}$
Sensitivity (Recall)	It is the true positive rate. It models the ability to predict true positive from all actual positive values.	$\frac{TP}{TP + FN}$
Specificity	It is the true negative rate. It models the ability to predict true negatives from all the actual negative values.	$\frac{TN}{TN + FP}$
F1 score	F1 score is the weighted average of precision and recall. The best F1 score is 1 and the worst F1 score is considered to be 0.	$\frac{2 * (Precision * recall)}{Precision + recall}$

4. FUTURE SCOPE

As the content on the internet will keep increasing exponentially in the future, demand for systems like Online news popularity prediction, which can help in improving business as well as quality of content produced, will also increase. The online news popularity prediction system can be enhanced further by choosing a better set of features for prediction, ones which contribute the most towards popularity of online news. Also, hyperparameter tuning can be performed on the machine learning models for improving the overall performance. The news content related features can be put more focus on, for improving the quality of news articles.

5. CONCLUSION

Choosing the most suitable techniques or algorithms for a problem is considered one of the most crucial steps for achieving a better performance. In machine learning domain, a number of techniques are available for solving any problem, depending on the type of problem at hand and the kind of data available for the model to train on, different machine learning algorithms give varying results. For effectively selecting the right tools and techniques, in this work, a comparative analysis of various machine learning techniques for the different stages of the machine learning process, starting from data pre-processing till model evaluation, was done for a particular problem of predicting the popularity of online news prior to its publication. It was observed that different approaches had their own benefits and drawbacks and hence should be chosen carefully based on one's requirements and needs.

REFERENCES

[1] A. Khan, G. Worah, M. Kothari, Y. H. Jadhav and A. V. Nimkar, "News Popularity Prediction with Ensemble Methods of Classification," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bangalore, 2018, pp. 1-6, doi: 10.1109/ICCCNT.2018.8494095.

[2] A. Chopra, A. Dimri and S. Rawat, "Comparative Analysis of Statistical Classifiers for Predicting News Popularity on Social Web," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Tamil Nadu, India, 2019, pp. 1-8, doi: 10.1109/ICCCI.2019.8822230.

[3] Y. Li, Q. Peng, Z. Sun, L. Fu and S. Khokhar, "A Two-stage Prediction Method of News Popularity only using Content Features," 2018 13th World Congress on Intelligent Control and Automation (WCICA), Changsha,

China, 2018, pp. 767-772, doi: 10.1109/WCICA.2018.8630557.

[4] F. Namous, A. Rodan and Y. Javed, "Online News Popularity Prediction," *2018 Fifth HCT Information Technology Trends (ITT)*, Dubai, United Arab Emirates, 2018, pp. 180-184, doi: 10.1109/CTIT.2018.8649529.

[5] Wei, W. U., D. U. Wencai, X. U. Hongzhou, Z. H. O. U. Hui, and H. U. A. N. G. Mengxing. "Popularity Prediction of Online News Based on Radial Basis Function Neural Networks with Factor Methodology." (2016).

[6] Liao, Dongliang, Jin Xu, Gongfu Li, Weijie Huang, Weiqing Liu, and Jing Li. "Popularity prediction on online articles with deep fusion of temporal process and content features." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 200-207. 2019.

[7] K. S. Naga Haritha, P. Vijaya Kumari, C. Nikhila Naga Jyothi, K. Manoj Kumar Naik4 "Predicting Online News Popularity", 2019 International Journal for Research in Applied Science & Engineering Technology (IJRASET)

[8] Rathord, Priyanka, Anurag Jain, and Chetan Agrawal. "A comprehensive review on online news popularity prediction using machine learning approach." trees 10.20 (2019): 50.

[9] D. Deshpande, "Prediction & Evaluation of Online News Popularity Using Machine Intelligence," *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, 2017, pp. 1-6, doi: 10.1109/ICCUBEA.2017.8463790.