Palestine Journal of Mathematics

Vol 14(Special Issue II)(2025) , 257–263

From Click to Predictions: Understanding Online News Popularity Through Machine Learning

Vellengara Adithi Sunil, Arshan Homi Dastur, G K Revathi and G Hannah Grace

MSC 2010 Classifications: Primary 57Z25; 62R07.

Keywords and phrases: Machine Learning, Online News Popularity, World Wide Web, Feature Selection, Mashable

The authors would like to thank the reviewers and editor for their constructive comments and valuable suggestions that improved the quality of our paper.

Corresponding Author: G K Revathi

Abstract *In a world dominated by the World Wide Web where information is easily and readily available to the people, online news has become an indispensable part of our daily lives to keep us updated of the things happening all around us. The number of times a news article is shared shows how popular the news is. Online news popularity prediction has become a popular research topic since it can help online news presenters, advertisers and writers to deliver highly competitive and reliable news to their viewers. This paper aims to find the best method as well as the set of features which predict the popularity of online news articles to the greatest extent, using machine learning algorithms. The data was acquired from Mashable, a very popular online news website. The main objective of this paper is to assess the performance of different algorithms on real world data. Feature selection techniques have been employed to decrease features and yield improved performance.*

1 Introduction

Online news popularity research is a cross-disciplinary field focused on unraveling the intricacies of how news content is disseminated and received in the digital era. Scholars within this realm delve into a myriad of factors influencing the popularity and virality of online news articles, encompassing social media's role, audience interaction, headline construction, content attributes, and the sway of algorithms. Through their exploration of these facets, researchers seek to illuminate the shifting landscape of news consumption, the consequences of digital platforms on public discourse, and the mechanisms governing information propagation in an increasingly interconnected and digitized global landscape. This research carries significant implications for journalism practices and our comprehension of the constantly evolving media environment.

Predicting the popularity of online news is a crucial and complex task in the modern digital information era. The significance of news, which delivers noteworthy daily events to the public, holds substantial importance for viewers and audiences [1]. Several studies have contributed to the field of online news popularity prediction, employing various machine learning techniques and methodologies. Machine learning techniques are used on the datasets to get precise prediction on the news popularity this will help media people to get an idea beforehand whether their article is going to be popular or not. Random Forest is the most widely used machine learning technique in online news popularity prediction. Random forest has been said one of the best machine learning techniques for popularity prediction, it has proved to give about 65 percent accuracy [6]. For doing this research we are considering a secondary data from a well-known online news website Mashable, there were research that were done on surveys and statistical measures applied to test hypotheses. However, relying solely on past data and statistical measures to predict news popularity has raised doubts about the accuracy of such predictions [4]. Data will be pre-processed first then we will be doing forward feature selection and backward

feature selection to narrow down the features that will be used for popularity analysis. Here we are focusing on Logistic Regression Model and Extremely Randomised Tree Classifier (Extra-Tree Classifier) with and without feature selection and compare them using the metrics accuracy. Then we will evaluate the model and find the strategies that will be helpful in predicting the popularity of the online news.

2 Objective

The main objective of this research is to address several key objectives. First, it seeks to test the accuracy of the predictions it makes, providing a reliable tool for assessing the likely popularity of news articles or content. Second, an integral aspect of this endeavor is identifying the crucial features that contribute to the online popularity of news pieces. By understanding these important factors, content creators and marketers can optimize their strategies. Furthermore, the model will shed light on the underlying factors that influence the popularity of news, helping media organizations and individuals tailor their content effectively. Ultimately, this research aims to bridge the gap between various studies and approaches, providing insights into the most effective algorithms for predicting online news popularity, thus contributing to the enhancement of content dissemination strategies in the digital age.

3 Methodology

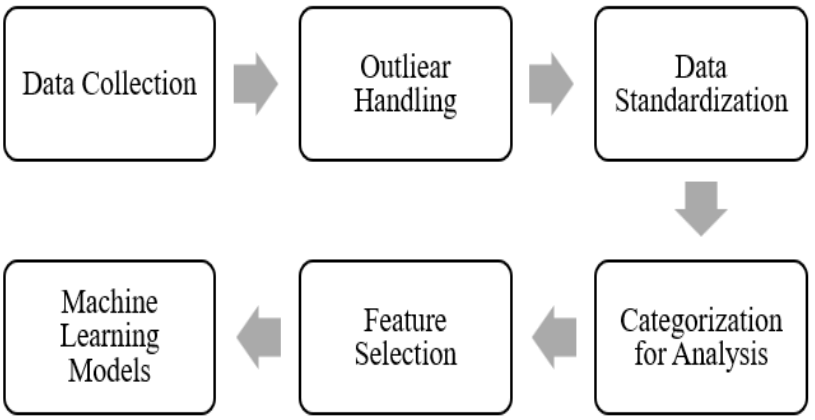

Figure 1. Flow chart of the overall methodology

The dataset comprises 39,643 news articles from the online news platform Mashable, collected over a two-year period spanning from January 2013 to January 2015 and sourced from the UCI Machine Learning Repository. Our initial data preprocessing involved thorough cleaning, encompassing the removal of missing values, erroneous data, and irrelevant columns. In particular, we eliminated "Url" and "Timedelta" since they had no predictive significance for the popularity of news articles. To assess variable relationships, we employed a heatmap (Figure 2) and computed a correlation matrix, allowing us to identify and eliminate highly correlated, redundant variables.

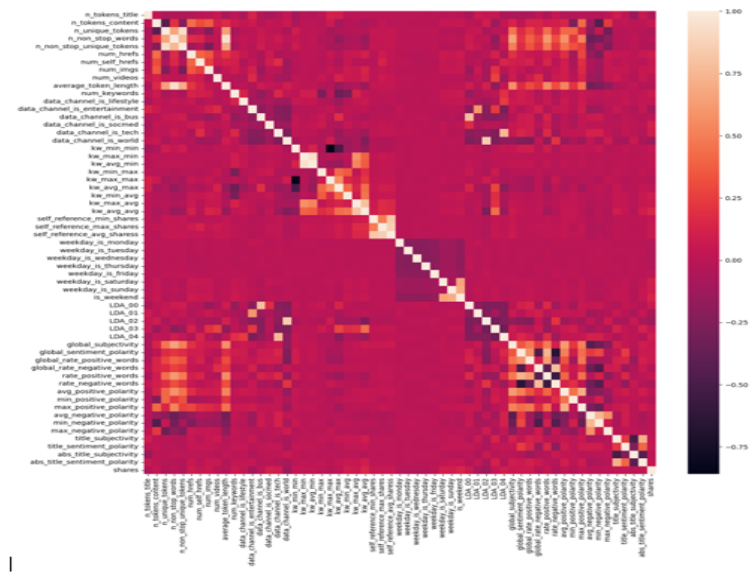

Figure 2. Heatmap showing correlation between the attributes

Ensuring data reliability, here we have addressed outliers by creating box and whisker plots for each variable. Using the interquartile range (I.Q.R) method, we identified values above Q3+(1.5I.Q. R) and below Q1-(1.5I.Q.R) as outliers and subsequently removed them from the dataset. Standardization was then applied to make the data suitable for analysis, transforming it to possess a mean of 0 and a variance of 1. We utilized the StandardScaler function from the pre-processing module in the Python sklearn library. To facilitate analysis, we categorized articles based on a threshold determined by the mean number of shares for each article. This threshold allowed us to encode data values as 0 (indicating unpopularity) or 1 (indicating popularity), simplifying the predictive modeling process.

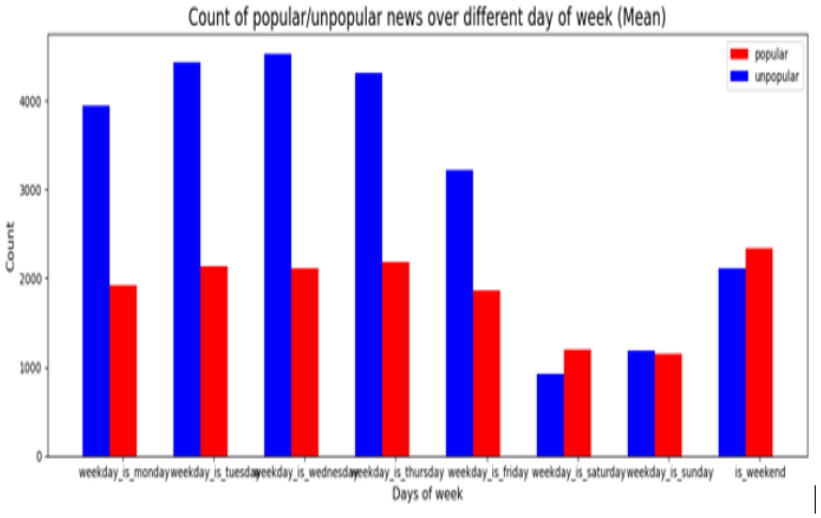

Figure 3. Histogram indicating the count of popular and unpopular news over different days of the week

Feature Selection: As part of data reduction, which aims to reduce the size of the dataset while maintaining crucial information, feature selection entails choosing a subset of pertinent

features from the dataset.For our dataset, we have implemented the forward feature selection technique. Forward selection is an incremental approach that begins with an empty feature set, progressively incorporating the feature that enhances the model's performance the most in each iteration until further additions fail to yield improvements in the model's performance. In this research project, we have initially applied the machine learning algorithms on the cleaned dataset (without removing any features), and then applied the same algorithms after performing forward selection. We have observed a noticeable difference in performance with and without feature selection. Another method that has been implied here is recursive feature elimination with cross validation. By successively deleting 0 to N features, the Recursive Feature Elimination with Cross-Validated (RFECV) technique systematically determines the best feature subset for a given estimator. In this method, the model is fitted multiple times, and its performance is evaluated using measures like accuracy, cross-validation scores, or ROC-AUC. The method keeps the subset of features that maximizes the selected measure while gradually removing other features. Recursive Feature Elimination (RFE) essentially increases the efficacy and efficiency of the model by repeatedly eliminating the least significant features.

Machine Learning Models Used: *In this research project, we have used Logistic Regression and ExtraTrees Classifier algorithm to perform popularity prediction.*

1.Logistic regression is primarily used for classification tasks in which the objective is to predict the likelihood that a given instance will belong to a given class or not. This statistical algorithm is a powerful tool for well-informed decision-making because it evaluates the relationship between a set of independent variables and a binary dependent variable.

2. The Extremely Randomized Trees Classifier, or Extra Trees Classifier, is an ensemble learning method that generates classification results by aggregating the outputs of multiple uncorrelated decision trees within a "forest." It closely resembles the Random Forest Classifier, with the key distinction lying in how the decision trees are constructed.

4 Related Works

Akyol, Sen (2019) discussed that people engage with news website by sharing news links and expressing their opinions on various platforms of social media. They have mainly focused upon the supervised learning techniques from machine learning to predict the popularity of news article. In this they have classified machine learning techniques into two phases. First a dataset is used for training the classifier algorithm, where the algorithm learns pattern and relationships within the data, Second, the performance of this pre-trained algorithm is tested on a separate dataset called the testing data to evaluate its predictive capabilities. For this study twelve datasets were collected, grouped into four categories. To accomplish this , three different machine learning algorithms were employed, Gradient Boosted Trees, Multi-layer Perceptron, Random Forest. To assess the performance of these algorithms mean square error, root mean squared error, and r squared value are used. These metrics help gauge how well the algorithms are performing in terms of accuracy and precision in predicting news popularity.

Hassija, Vikas (2023) detailed upon the forecast was supported by empirical data gathered via surveys and statistical analyses used to evaluate theories. Here, they have listed three elements that are essential to developing a thorough prediction model. The model can gather a lot of information about every news article by combining text, image, and meta features. Using the Root Mean Squared Logarithmic Error (RMSLE) as a performance metric, the proposed model's overall performance is assessed. It also seeks to pinpoint the salient characteristics. This analysis helps determining the extent to which the prediction model relies on text and image features, providing insights into the relative importance of each feature type in predicting news popularity.

Deshpande, D. (2017) stated that online news popularity is determined by various factors such as the number of comments from visitors, shares on social media and likes on news articles. Initially, the dimensionality of the data is decreased using a method known as Latent Dirichlet Allocation (LDA).This step helps streamline the features used for prediction. Subsequently, three different machine learning algorithms are used here for prediction. AdaBoost, LPBoost, and Random Forest. These algorithms are used to make predictions about news popularity based on the reduced features set. To evaluate the performance the assess the effectiveness of the models using various evaluation measures, typically include accuracy and F- measure. Among the three methodologies tested, the Adaptive Boosting(AdaBoost) emerges as the best model for predict-

ing news popularity. AdaBoost achieves an accuracy rate of 69 percent and an F-measure of 73 percent , indicating its effectiveness in this task.
Hensinger,Elena,Flaounas,Christiani (2013) elaborated on the viewpoint of a competitive scenario where the articles that are most popular are the ones that appealed to readers the most on that specific day. The training set, composed of pairs of documents—one popular and the other appearing on the same date and page but not gaining popularity is used to determine the parameters of this linear function. They demonstrated how their approach can identify articles with the greatest potential for popularity and identify the keywords that have the biggest impact on appeal function.
Namous, F., Rodan, A., and Javed, Y. (2018)
mentioned that in the online news popularity prediction the two best performing models identified in the research were Random forest and Neural Network. These models achieved an accuracy rate of 65 percentage when configured with optimal parameters. This level of accuracy can be valuable for online news companies as it allows them predicting the popularity of news articles before they are published. The study involved a comparison with previous research conducted on the same dataset, further validating the effectiveness of their approach. They aimed to develop a predictive model for online news popularity, and the results suggest that Random forest and neural network are the most suitable algorithms for this task. News popularity prediction helps the publishers get an judgement about publishing an article by checking its predicted popularity.
Guan,Peng, Li,Zhu (2017) stated the challenges associated with predicting the popularity of news articles where the abundance of information and intense competition for user attention. The main focus is on information about user interactions or historical data. Here the hierarchical neural network model is proposed. This neural network is designed to create distributed representations on news articles, allowing it to capture specific sequence of words or phrases and model the sequential relationships between sentences within articles. The goal is to leverage the article's content to make predictions about their popularity. The prediction was done on real data from an online news portal that proposed methods outperforms traditional methods for popularity predictions. Notably, the model excels in identifying articles that are likely to become popular, even when only content-related features are considered addressing the cold start problem effectively.
Khan, Aasim, Worah, Jadhav, Nimkar (2018) pointed out that online articles are primary means of spreading information, and understanding what types of articles will be more popular is crucial for media outlets. The approach is based on Random Forest machine learning algorithm. Random Forest is a model that can make predictions by combining the results of multiple decision trees. However, since processing a large amount of data can be computationally intensive, they have employed dimension reduction techniques, specifically principal component analysis (PCA), which lowers the data's complexity. The performance the models are accessed by ROC area values, which are common metric for evaluating the accuracy of classification models. The results indicate that Random Forest based approach outperforms other models such as Classification and Regression Trees and 4.5 algorithm in predicting article popularity. This suggests that Random Forest is a viable technique for precisely predicting the popularity of online publications, with the help of dimension reduction by PCA.

5 Discussion

Here the utilization of machine learning techniques, specifically Extra Tree Classifier and Logistic Regression, in conjunction with feature selection methods like Forward Feature Selection and Recursive Feature Elimination with Cross-Validation (RFECV), provides valuable insights into the factors influencing the online news articles' popularity. The effectiveness of these techniques was evaluated by analyzing data from the well-known online news source Mashable. The outcomes demonstrated how well the extratree classifier captured intricate linkages and distinguished important characteristics associated with news popularity. Furthermore, the performance of the algorithms was enhanced by forward feature selection and recursive feature elimination with cross-validation, which helped choose pertinent features and lower the dimensionality of the dataset. We delve into the comparative analysis of the two feature selection methods, exploring their impact on model performance, interpretability, and computational efficiency. It is crucial to assess how these methods complement each other and whether there are

specific scenarios where one outperforms the other, considering the strengths and weaknesses of both Extra Tree Classifier and Logistic Regression models.

6 Results

The metric which we have used to evaluate the model performance here is accuracy.

Table 1. The accuracy results, specifying the algorithm used and the corresponding feature selection method applied

Algorithm	Feature Selection Method	Accuracy
Logistic Regression	None	0.5740195612952236
ExtraTrees Classifier	None	0.6489412211565854
Logistic Regression	Forward Feature Selection	0.6354572215364163
ExtraTrees Classifier	Forward Feature Selection	0.6164656727756148
Logistic Regression	RFECV	0.5730699838571836
ExtraTrees Classifier	RFECV	0.6450479536606211

7 Conclusion

The paper aims to contribute to the enhancement of content dissemination strategies in the digital age by improving the accuracy of predictions, understanding the critical features, and shedding light on the factors that drive news popularity. The results show that ExtraTrees Classifier performs better than Logistic Regression without feature selection, and feature selection can have a positive impact on predictive accuracy. While forward feature selection performed less well with the ExtraTrees Classifier, the Recursive Feature Elimination with Cross-Validation technique continuously improved the model accuracy. This underscores the notable contrast in performance observed when employing different machine learning algorithms.

References

[1] *Akyol Kemal, and Baha Şen.* "Modeling and Predicting of NThews Popularity in Social Media Sources, *Computers, Materials and Continua 61–1 (2019).*

[2] *Deshpande, Dhanashree.* Prediction and evaluation of online news popularity using machine intelligence, *International Conference on Computing, Communication, Control and Automation (ICCUBEA). IEEE, (2017).*

[3] *Guan, X., Peng, Q., Li, Y., and Zhu, Z. (2017, October).* Hierarchical neural network for online news popularity prediction, *Chinese Automation Congress (CAC) 3005–3009. IEEE (2017).*

[4] *Hassija, V., Arora, A., Bansal, S., Yadav, S., Chamola, V., and Hussain, A.* A Novel Multimodal Online News Popularity Prediction Model based on Ensemble Learning *(2023).*

[5] *Hensinger, E., Flaounas, I., and Cristianini, N.* Modelling and predicting news popularity, *Pattern Analysis and Applications,* **16**, *623–635 (2013).*

[6] *Namous, Feras, Ali Rodan, and Yasir Javed.* Online news popularity prediction, *Fifth HCT Information Technology Trends (ITT). IEEE, (2018).*

[7] *Khan, A., Worah, G., Kothari, M., Jadhav, Y. H. and Nimkar, A. V.* News popularity prediction with ensemble methods of classification, *9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (1–6). IEEE, (2018).*

Author information

Vellengara Adithi Sunil, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Chennai, India.
E-mail: vellengara.adithi2023@vitstudent.ac.in

Arshan Homi Dastur, Department of Mathematics, School of Advanced Sciences Vellore Institute of Technology, Chennai, India.
E-mail: jarshanhomi.dastur2023@vitstudent.ac.in

G K Revathi, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Chennai, India.
E-mail: gk_revathi@yahoo.co.in

G Hannah Grace, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Chennai, India.
E-mail: hannahgrace.g@vit.ac.in