An adaptable approach in using Machine learning towards predicting the Popularity of news articles.

Rohit Sekhar Kondajogi
Symbiosis Institute of Business Management, Symbiosis International (Deemed University), Bengaluru India

ABSTRACT

The vast and increased utilization of the web and the arrival of the data technology field has led to a new age where individuals are starting to read online news frequently. Hence, online news has become the fundamental source of information for most individuals, and anticipating the prominence of online news has become a discussed issue that can't be neglected. It could help in assisting authors with introducing serious and highly readable news. Number of shares an article gets is considered as one of the most obvious factors in determining its popularity. In this paper, we apply distinctive machine learning methods to anticipate the quantity of shares and categorize them as well known and unpopular. The information has been assembled from the UCI machine learning repository from Mashable. Linear regression and classification techniques like decision tree, SVM, and logistic regression are utilized to the data set. The performance of these methods is measured by their accuracy, precision, and recall measures. The aim is to find a reliable model with a prediction accuracy of 70 %. Having found a reliable prediction model, this work can then be used by online news agencies to anticipate their popularity based on content and to make changes accordingly and help news agencies in adopting promising advertising strategies.

Keywords
Machine learning, Linear regression, Classification techniques, Decision tree, SVM, Logistic regression.

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

According to author Obiedat, (2020), an article is a composed work that can be distributed either in printed copy format (e.g. Newspapers) or soft copy format (for example online article). Due to the ever-developing utilization of the web, social networks and technological revolution (e.g. the use of smartphones) thousands of online news sites were accessible to online readers. So users tend to peruse online news more than papers, which makes online news the primary source of data for the significant piece of the network. An article is seen as a well-known piece of news if it is among the most read and engaging articles on a specific day of distribution in a certain news outlet. The popularity of an article can be estimated dependent on a few variables, for example, the quantity of views, comments, likes, votes, or shares through social networks or by email as proposed.

The fame of news is of great significance and value to many sectors, like a business, marketing and online advertising, and even in political activities, since individuals lean tend to mostly lean towards reading the most mainstream articles and offering it to companions. This would probably influence various organizations on the planet spend up to 30% of their working capital on online marketing, therefore, online news sites and social media pages content to pull in more readers and attempt to improve the nature of online articles before their distribution. Moreover, news sites can employ predictions to highlight their popular news, and arrange it on their home page accordingly, to pull in the readers by identifying their interest, furthermore concentrating on the significant and connecting with news that they will discover intriguing. Thus, online news sites can designate their assets better to compose stories on the chosen subjects at the correct time. Furthermore, online readers can filter the huge amount of available information quickly and easily, and centre on the most significant ones. Then again, it can help governments apportion unsafe news and stop distributing such news.

Online advertisement strategies have become progressively keen on understanding reader behaviour and anticipating the articles that may increase large user notice. Consequently, it is vital for social media sites and online news authors to construct an automated model that anticipate the fame of news prior to their distribution and such models can be implemented through the use of business intelligence and data mining tool.

Foreseeing the fame of online news is a challenging and complex assignment for some reasons since such a large number of components influence the notoriety of a point. To begin with, it is difficult to gauge the nature of the substance or its pertinence to the readers' interest, other than the unavailability of the context outside the web, as well as the local and geographical conditions that may influence the population and make the prediction more difficult. Also, the popularity depends totally on user behaviour, interest, feelings, and point of view which is very hard to be predicted. Fame depends absolutely on user behaviour, interest, feelings, and point of view which is exceptionally difficult to be anticipated as written in a paper in 2001 by Chan, C. H., Sun, A., & LIM, E. P. Prediction, particularly preceding distribution, needs numerous biased highlights. Moreover, it is difficult to examine the content of the web and semantic information, and the structure of the network. The connection between the various layers of the web present another challenge and the page structure intricacy like the first-page location and the subsequent page make the prediction considerably troublesom .

Many authors who have explored the scope of online news popularity believe that supervised learning/Predictive models present a fitting technique. All the papers have stressed on a classification approach since it can classify the data into predefined paired classes. In this paper we try to

build a model based on machine learning to anticipate the fame of online news before its distribution by utilizing regression and classification techniques. The objective is to get a model with high predictive power which decision-makers can rely upon. Binary classification is used to classify an article as famous and unknown depending on the count of the article's shares across social media sites.

For this research two supervised machine learning techniques have been used which are linear regression and classification. In this paper, we attempt to predict the quantity of shares a news article would get on publishing and its' influencing factors by fitting a linear regression model. Logistic regression is used as the main classification technique. It is employed to create a model to foresee the count of shares and also to classify a news article as popular and unpopular. This is done using binary classification. From these models, one can get both qualitative and quantitative insights about factors influencing the popularity of news articles.

Literature review

Predicting online news popularity is a hot topic and many scientists have researched in this field. Most of them deal with text classification which mainly talks about techniques like SVM, random forest and so on. In the paper by Jotikabukkana, P., Sornlertlamvanich, V., Manabu, O., & Haruechaiyasak, C. in 2015, a reliable method to classify the content on social media by using online news is proposed . Along with the commonly known Term frequency reverse document frequency weighting technique (TF-IDF), Word Article Matrix (WAM) is utilized as guideline techniques in this investigation. They analyse the times series of tweets with explicit period to demonstrate their method which can separate catchphrases, categorize social media text productively, and can illustrate the evolution of social behaviour on a happening. Twitter restricts the tweet length to 140 characters. In "Automated online news classification with personalization. (2001)" depict a working news classification framework, called Categoriser which classifies online news. Categoriser uses SVM technique to classify news stories. They organized Categoriser to classify reports which are mainly financial news from the Channel News Asia. As a rule classification, they received an established arrangement of classes from the Reuters assemblage. The Reuter's assemblage was picked considering the way that its classifications are immovably related to financial institutions and budgetary perspectives. In a 2010 paper, Krishnalal G, Babu S Rengarajan and K G Srinivasagan Proposed a machine learning model for internet news classification based on Hidden Markov Model (HMM) and Support Vector Machine (SVM). Popular media such as The Hindu, The New Indian Express, Times of India, Business Line, and The Economic Times served as source for the data set. The proposed model which a hybrid is of HMM and SVM methods proves to be reliable with excellent results. Ikonomakis, Emmanouil & Kotsiantis, Sotiris & Tampakas, V. Explained the text classification process through machine learning techniques. Various stages namely pre-processing, feature extraction, feature selection and modelling were describing in detail by them. Popular methods like Naïve Bayes, Random Forrest, SVM and evaluating parameters for these methods were discussed exclusively.

Obiedat.R Concentrated on utilizing techniques to assess and look at the performance of five classification models which are Random Forest, Bayes Net, Logistic Function, Simple Cart and C4.5 on the online news dataset, He Tried to categorize an article as famous or unknown, depending on the number of shares, where the threshold is set to 1,400. The performance of these five classification models has been assessed utilizing a few of the generally normal and famous evaluation metrics in the data mining field. Results show that Random Forest was the best classifier compared with other introduced models since it has the most noteworthy Accuracy estimation of 66.7869% and 0.3297 Kappa Statistics and the most reduced RMSE estimation of 0.4586. Subsequently, the Random Forest could be very much helpful for online news popularity prediction. Balali, A., Asadpour, M., Faili, H., Balali, A., Asadpour, M., & Faili, H. in their paper Identified the features of an article that urge individuals to leave a comment for it. They proposed a machine learning approach to predict the volume of comments utilizing the data that is extricated about the users' activities on the site pages of news organizations. The datasets that were used in this paper were gathered from a few online news agencies in Iran. They daily publish news stories about recent developments in Iran and abroad in various classifications, for example, governmental issues, economy, culture, society, innovation, sport, and so forth. In this paper, the authors additionally classified a few features for weighting words based on users' interests that can be valuable in concocting devising slogans for electoral campaigns and advertisements. Kathal, A., & Namdev, M Examined the execution of various machine learning algorithms such as Hybrid SVM-RF, AdaBoost, LPBoost, and KNN are implemented to the news popularity. However, Hybrid SVM-RF ended up being the best model with an accuracy of 99.6% for binary classification. Van Canneyt, S., Leroux, P., Dhoedt, B., & Demeester, T. in a paper written in 2017 built a model to investigate patterns of web-based news viewing. Post an exhaustive examination of alike designs, they proved that thoroughly selected fundamental behaviour lead to appropriate models, and also described the impact of daytime versus night on the overall view patterns. They confidently proved that the use of variables related to the substance, and materialistic behaviour improve prediction accuracy significantly, contrasted to existing methodologies which consider only those factors directly related to the popularity of articles. They focused on the gradient boosting tree method rather than regression. Haritha, K. S. N in a 2019 paper Explored different techniques like Logistic regression, Random Forest, Adaboost algorithm to build a system to foresee the fame of a news article using features like the number of shares, comments, likes etc. Akyol, K., and Şen, B in their 2019 paper used supervised learning methods like Gradient Boosted Trees, Multi-Layer Perception and Random Forest to predict news popularity in social media networks. In a 2012 paper regarding forecasting populairty by Bandari, R., Asur, S., & Huberman, B, both regression and classification algorithms are used to forecast ranges of fame on twitter with an 84% accuracy. Linear regression, SVM and KNN algorithms have been used by authors for their purpose.

Orellana-Rodriguez C., & Keane M. T in their paper 'Modelling and Predicting News Consumption on Twitter' focused on examining the dynamics in news on Twitter. They were inspired to find what drives readers to consume news, and hence create a consumption prediction model. The authors developed Twitter News Model (TNM). They concluded that news inspirations, trailed by news perspectives and news convictions, impact users' behaviour of news consumption on Twitter.

After a critical analysis of the literature review of existing studies, it is understood that many kinds of research in the past have worked upon evaluating the performance of different algorithms to predict popularity. There was no significant work to identify influencing factors. However, this paper proposes a legitimate linear model to predict the count of shares a specific article would get and the variables influencing it and to explain how to test goodness of fit of a model once it is implemented. This paper focuses on identifying influencing factors and exploring how well such factors foresee and predict the fame of news articles. The objective here is to identify factors which have the greatest impact on the popularity and apply various supervised learning models to predict siad popularity. The model with the best accuracy will be a reliable one which can be used by several online news agencies for reaching maximum visibility which could news outlets help in streamlining the different types of news.

Methodology

Original set

The data set i.e. secondary is collected from the UCI machine learning repository published by Mashable. It consists of a total of 39644 articles and 59 attributes. However not all attributes are used as predictors, some are eliminated as they don't contribute significantly for the model. The analysis has been done in R starting from cleaning, through exploratory analysis and finally predictive models.

Machine learning algorithms

Linear regression

This is one of the simplest supervised learning techniques which is used to predict an outcome in numerical value using predictors. This method tells about the most significant influencing factors for the outcome variable. The impact of the outcome variable can be determined by the coefficients of predictors from the model. In this way, this technique can provide useful insights which can be worked upon later. The dependent variable is count of shares a specific article acquires. Predictors are count of tokens in the title and content, count of videos and images, number of links, genre (channel) of the article, publishing day. However, these are the selected features from 59 attributes which will significantly predict the popularity (shares).

Multilinear regression is used to find a relationship between a target variable and predictors by fitting a linear model based on the train set of original data. Mathematically written as 'Target =Residual +Fit' in which 'Fit' term indicates the expression $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_n X_n$. The "Residual" term indicates the fluctuations of the noted values 'y' from their means '$\mu_{y'}$, which are normally distributed with mean 0 and variance 'σ'. Below is the expression for linear regression with multiple predictors

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_n X_n$.

Where Y represents the outcome variable and β_0 is intercept or the constant term. $\beta_{1,2,n}$ represents coefficients of predictors ($X_1, X_2 .. X_n$).

Logistic Regression

This is a classification technique which is also one of the most widely used simplest methods in the category. This paper proposes a binary classification model which can predict whether a particular article would be popular or not. The predictors taken for this method is the same as that of the linear regression model. However, a new column has been added which assigns binary values to popular and unpopular articles based on the benchmark set (shares = 1600). The benchmark has arrived after critical analysis of the literature review of the work by Obeidat, Chee-Hong, Ren, H. and Yang, Q. The outcome variable is expressed as either 0 or 1. '0' indicates not popular and '1' indicates popular. The model gives the probability of the occurrence of a particular class.

The logistic regression model is expressed as below

$P(X) = e^{(\beta_0+\beta_1 X)} / (1 + e^{(\beta_0+\beta_1 X)})$

Where P (X) is the probability, β represents coefficients of predictors and X denotes the predictors or variables. Taking the natural log (ln ()) on both sides gives the below equation.

$$\ln\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X$$

$\frac{P(X)}{1-P(X)}$ Is called odds ratio hence this model gives log-odds using predictors which is then changed to the odds ratio by removing ln() from both sides. So logistic regression predicts maximum likelihood (odds ratio) of occurrence for a particular class.

Decision tree

Decision tree classifier algorithm is applied, with the objective to classify popular and unpopular news article. Features used in the algorithm are the same as that of logistic regression. The decision is taken depending on the rules developed by the algorithm. It serves as a base for other supervised algorithms. From the decision tree, It can be understood clearly what factors are considered while deciding whether an article is popular or not.

Support Vector Machine

SVM is one of the most common and frequently used machine learning algorithms for both classifications as well as for regression questions. Nonetheless, it is predominantly used for classification problems as a better alternative for simple logistic regression. In this paper linear SVM have been used to model the data. The features include all variables except, average token length, average, maximum, and minimum shares of referenced articles. These are

selected from exploratory data analysis after understanding the impact of such variables on popularity.
The performance of these models is checked via accuracy, precision and recall values.

Feature selection

Feature selection is a way of choosing only those variables/factors from the original data set that can model the data in the best possible way. The result is the variable data set which is relevant for the construction of models. The factors for different algorithms are chosen using exploratory data analysis after understanding the impact of such factors on the outcome variables shares/popularity. Only those variables which significantly affect the outcome variable via visualization in bar plots and scatter plots. This analysis provided a broad view of the behaviour of different factors against 'shares'. Detailed analysis with graphs is shown in a later section.
The reason for selecting these features specifically was because accuracy is improved when the right subset is chosen. feature selection also preserves original features so that you can go back and see which features are important.

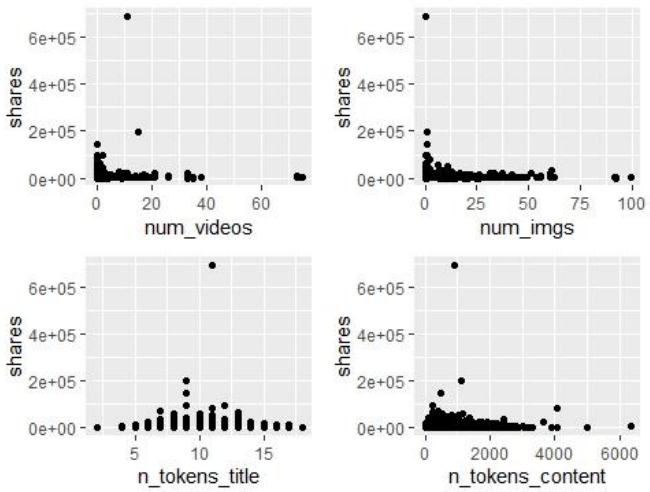
Table 1: Selected features

Word	Count of words including the title and content
Links	Count of links
Media	Count of images and videos
Time	Day of week
Theme	Genres (Business, Entertainment, Social media, World, Technology)
Target	Shares/Popularity

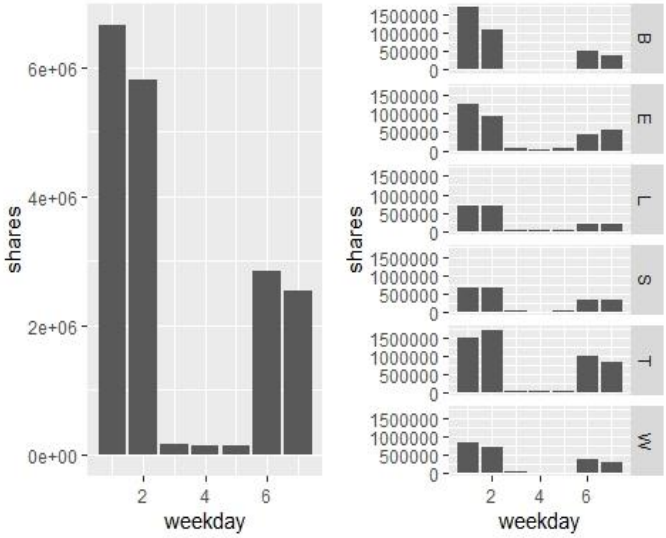
Results

Exploratory Data Analysis

Basic plots such as scatter plots and bar plots are utilized to explore the data set and to understand critical factors. The factors which have a significant correlation with outcome variable 'shares/popularity'. Below are the graphs plotted for the selected factors to be used for modelling. From the visualization, these factors seem to critically affect 'shares/popularity'.



It can be observed that the majority of the data points is concentrated between 0 and 40 approximately and beyond 50 points seem to be dispersed far from each other. Articles having number of images between 1 and 30 seem to be more popular than the rest. It seems to be that the number of videos is between 0 and 15 for articles to be popular. The number of words in the title seems to be uniformly distributed with the majority of the data points lying between 7 and 12 approximately. Such articles seem to be more popular than the rest. Most of the data points are concentrated between 0 and 2000 beyond which no of shares decreases as the points get dispersed out.
In the below graph, row corresponds to weekdays (1= Monday,7 = Sunday) and the letters B, E, L, S, T, W represents genres of the articles as Business, Entertainment, Sports, Tech, and World respectively. The plot provides insights about the potential genres and weekdays as follows. Articles published on Monday seem to have the maximum number of shares followed by Tuesday, Saturday and Sunday. During mid of the week, articles don't get shared frequently which is quite strange. From the below plot, It can be seen that best articles with the highest share popularity belong to the genre "Business " and "Entertainment" followed by "World " and "Tech". Best articles are often published on Mondays and Tuesdays.



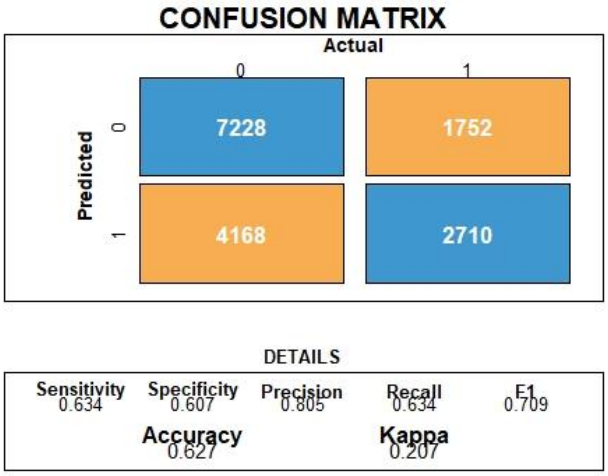
Machine learning algorithms

As stated earlier, four models were created by making use of the training set and run on the test set. These models were analysed to check and compare their accuracy, precision and recall via confusion matrix. The benchmark of 1600 shares was set as differentiating criteria after a thorough analysis of existing studies [5, 6, 8, and 10].

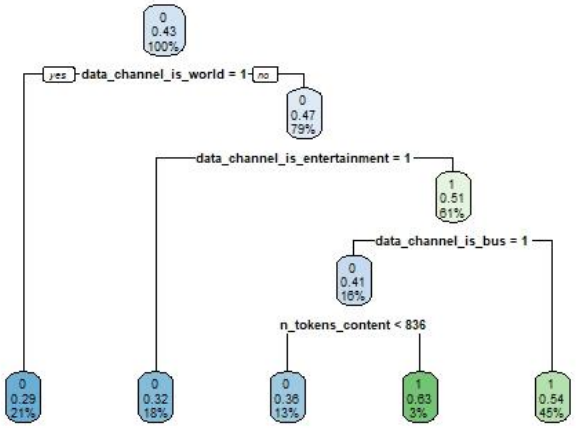
Logistic regression

Evidently, from the confusion matrix, this model has an accuracy of 63 % with set ratio 60:40. The accuracy improved from 60 % accuracy to 63 % when the training set decreased from 80 % to 60 % of the total data set. Precision tells that the model identified 80.5% values as relevant ones

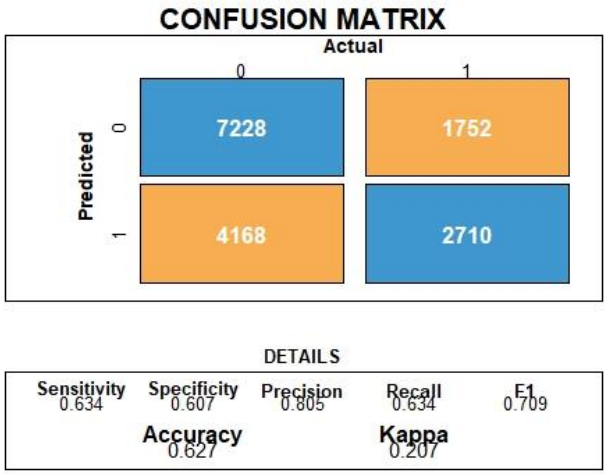
(class 1 of test set) and recall tells the model correctly classified 63.4% of the relevant results.



Decision tree



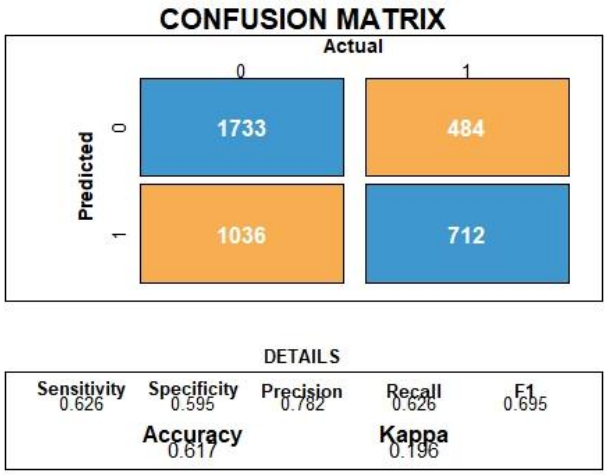
With 85 % as the training set, five rules are generated from the tree and can be understood that 'Business', 'Entertainment', and 'World' are the most promising genres to gain popularity. According to the tree, the content of the article should have at least 836 words to attain visibility among readers. The confusion matrix is the same as that of logistic regression with accuracy as 62.7%.



Linear Regression

This model was created with 90 % data as a training set. Accuracy improved as more and more data points were used in the training set. Although the model accuracy is very low close to 10.4 %.The variable with greatest impact on popularity is 'count of words' in the heading as per this model. Variables 'weekday' and 'count of words in content' seem to be not significant in the model hence can be ignored.

Support Vector Machine



This model was created using the SVM linear kernel technique. An accuracy of 61.7 % was achieved using a training set with 90 % of the data. Although lesser than the logistic model, the precision and recall are quite appreciable. This model has correctly classified 62.6 % (recall) of relevant points from 78.2 % (precision) of correctly identified relevant ones.

Results and Discussion

Table 2: Summary of results

Algorithm	Accuracy	R^2/Recall	Adj R^2/Precision
Linear Regression	10.4 %	10.1%	9.76%
Logistic Regression	62.7%	63.4%	80.5%
Decision Tree	62.7%	63.4%	80.5%
SVM	61.7%	62.6%	78.2%

According to accuracy, logistic regression model seems to perform better than all but the difference between precision and recall is higher than SVM and decision tree. On the other hand, the SVM model can classify 62.6 % of the data correctly from 78.2 % of data correctly returned by the algorithm despite having a little lesser accuracy than the logistic model. Hence the SVM model seems to be more reliable in this classification problem. Linear regression is the worst performing model in foreseeing the count of shares for a specific article. Although having such a poor accuracy (10.4 %), this model can roughly tell which the factors that are significant in predicting popularity are. It gives a fair idea about the factor which has the greatest impact on popularity. simpler models generally produce more precise predictions. Given several models with similar explanatory ability, the simplest is most likely to be the best choice. Start simple, and only make the model more complex as needed. The more complex you make your model, the more likely it is that you are tailoring the model to your dataset specifically, and generalizability suffers. A closer look into the decision tree provides better insights than a linear regression model.Also, It is clear that only three of the genres 'Business', 'Entertainment' and 'world' were considered at decision points irrespective of the training set, unlike linear model where all the genres were significant. The prediction performance of the decision tree is the same as that of the logistic model however it can be improved by adjusting the ratio of train and test data.

Conclusion

News popularity prediction is becoming a not to be ignored topic among many data scientists and analysts working in various agencies. The main reason for this era is a huge penetration of smartphones, web, and online news. Online news/blogs have become a major source of information for a majority of people around the world due to internet penetration which led to increased dependency. This paper focused on using the procedures to evaluate and compare the performance of four supervised learning algorithms which are linear regression, logistic regression, decision tree, and support vector machine on the online news data. Three classifications and one regression technique are applied in this paper. The classification models attempt to categorize popular and unpopular news from the data set depending on the count of shares where the threshold is set at 1600. The performance of these four models has been evaluated by the most common and popular evaluation metrics namely accuracy, precision and recall. Results show that logistic regression is the best classifier model as it has the highest accuracy of 62.7%. However, according to all the metrics, SVM appears to be more reliable classifier model as it has comparable recall (62.6%), precision (78.2%) and an appreciable accuracy. Before applying algorithms, an exploratory analysis was done to study the data set in depth and draw some interesting insights that could serve as a base for modelling. This also helped in selecting factors for modelling. The exploratory analysis helped in selecting the best features for modelling. This increased the speed of execution of algorithm which in turn led to better accuracy.
The popularity of news is significant and valuable for many sectors, like a business, marketing and online advertising, recommendation systems, and even in political activities, since individuals lean toward reading the most mainstream articles and offering it to companions. This would probably impact open intrigue and assessments. Moreover, news sites can employ predictions to highlight their popular news, and arrange it on their home page accordingly, to pull in the readers by identifying their interest, furthermore concentrating on the significant and connecting with news that they will discover intriguing. Thus, online news sites can designate their assets better to compose stories on the chosen subjects at the correct time. Furthermore, online readers can filter the huge amount of available information quickly and easily, and centre on the most significant ones. This can help governments apportion unsafe news and stop distributing such news to the public.

Future work

As observed from the results, despite having state of the art data set from Mashable no algorithm had reached 70% accuracy. Linear regression came out to be the worst model. One of the reasons could be model selection though it should be noted that feature selection must be heavily worked upon before going to the actual modelling. Although the original data set appeared to be a regression problem, linear regression did not work well as the data itself was not linear. Hence an appropriate method should be used to model such data. It is to be noted that as far as the scope of this study is considered, the substance of the articles hasn't been completely investigated. Many features related to content like LDA measures, polarity measures were removed as they didn't correlate significantly with popularity/shares as compared to other features like the number of images, several videos etc. News articles must be directly extracted so that text mining can be devoted to the data and features related to content are chosen appropriately. Thus, different models can be made by considering what the substance (content) discusses and this methodology could improve accuracy.

References

[1] Jotikabukkana, P., Sornlertlamvanich, V., Manabu, O., & Haruechaiyasak, C. (2015). Effectiveness of social media text classification by utilizing the online news

category. 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA).

[2] Chan, C. H., Sun, A., & LIM, E. P. (2001). Automated online news classification with personalization.4th International Conference on Asian Digital Libraries. Research Collection School of Information Systems.

[3] Krishnalal G, Babu S Rengarajan and K G Srinivasagan. Article: A New Text Mining Approach Based on HMM-SVM for Web News Classification. International Journal of Computer Applications 1(1):98–104, February 2010. Published By Foundation of Computer.

[4] Ikonomakis, Emmanouil & Kotsiantis, Sotiris & Tampakas, V.. (2005). Text Classification Using Machine Learning Techniques. WSEAS transactions on computers. 4. 966-974.

[5] Obiedat, R. (2020). PREDICTING THE POPULARITY OF ONLINE NEWS USING CLASSIFICATION METHODS WITH FEATURE FILTERING TECHNIQUES. Journal of Theoretical and Applied Information Technology, 98, 8

[6] Balali, A., Asadpour, M., Faili, H., Balali, A., Asadpour, M., & Faili, H. (2017). A Supervised Method to Predict the Popularity of News Articles. Computación y Sistemas, 21(4), 703–716.

[7] Kathal, A., & Namdev, M. (2018). Correlation Enhanced Machine Learning Approach based Online News Popularity Prediction. SMART MOVES JOURNAL IJOSCIENCE, 4(3),

[8] Van Canneyt, S., Leroux, P., Dhoedt, B., & Demeester, T. (2017). Modelling and predicting the popularity of online news based on temporal and content-related features. Multimedia Tools and Applications, 77(1), 1409–1436.

[9] Haritha, K. S. N. (2019). Predicting Online News Popularity. International Journal for Research in Applied Science and Engineering Technology, 7(4), 3451–3455.

[10] Akyol, K., Şen, B. (2019). Modelling and Predicting of News Popularity in Social Media Sources. CMC-Computers, Materials & Continua, 61(1), 69–80.

[11] Bandari, R., Asur, S., & Huberman, B. (2012.). The Pulse of News in Social Media: Forecasting Popularity. arXiv preprint arXiv:1202.0332.

[12] Orellana-Rodriguez, C., & Keane, M. T. (2018,July). Modelling and Predicting News Consumption on Twitter. Proceedings of the 26th Conference on User Modelling, Adaptation and Personalization.

[13] K. Fernandes, P. Vinagre and P. Cortez. (2015) A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

[14] Ren, H., & Yang, Q. (2015). Predicting and Evaluating the Popularity of Online News. Standford University Machine Learning Report.

[15] Karabulut, E. M., Özel, S. A., & İbrikçi, T. (2012). A comparative study on the effect of feature selection on classification accuracy. Procedia Technology, 1, 323–327.

[16] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[18] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

[19] Simkin, M. V., & Roychowdhury, V. P. (2008). A theory of web traffic. EPL (Europhysics Letters), 82(2), 28006.

[20] Kempe, D., Kleinberg, J., & Tardos, É. (2003, August). Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 137-146).