Online News Popularity Prediction

1st Alka Simon
CSE Department
PES University
Bangalore,India
alkasimon23@gmail.com

2nd K Vennela
CSE Department
PES University
Bangalore,India
kvennela1998@gmail.com

3rd Yeshaswini M
CSE Department
PES University
Bangalore,India
yeshasu555@gmail.com

***Abstract*—Online platforms publish hundreds of articles every day.The number of shares or comments that a particular article gets decides how popular that particular article is.So in this project, we aimed to predict the popularity of an article (by estimating the number of shares a particular news article can get) based on its relevant features ,that we could extract from the dataset provided by Mashable.We have modelled various models ranging from regressions to random forest and SVMs and compared those models.We have used step wise feature selection for dimensionality reduction. After predicting this,we recommend the changes that one needs to make to the present article before publishing it so that it can become popular.**

I. Introduction

Consuming news articles is an integral part of daily life.A widespread use of smartphones and the social networking sites has introduced people to the online news.Because of which there has been a growing interest in online news, which allow an easy and fast spread of information.These are of small size, have short lifespan and are of low cost. These properties has made them more effective.It can capture attention of internet users in a short period of time. Reading, sharing, commenting on these has become the part of peoples entertainment lives.

The number of comments and shares on an articles is of great importance which shows peoples interest on it. Using this we can also say that which news article has influenced public.Therefore, it would be of great use if can predict the popularity of the news article. The subject of popularity domain has been greatly used for home-page rankings for news articles in sites like yahoo etc.Popularity of an article can be determined by tracing the number of comments on it or the number of shares of the article.

Popularity prediction is a challenging task because of its difculty to measure the quality of content and how relevant the content is with respect to the interest of the users, prediction difculty of complex online interactions and information cascades ,local and geographical conditions,inaccessibility of context outside the web; social network properties and so on.

Predicting online news popularity is found to be useful in many fields.By using the advantage of predicting the popularity of a news article,any news organization or any blog publication organization ,even activists/politicians (e.g., to understand or influence public opinion) can have a better knowledge and understanding about different types of news articles that a particular user or viewer is interested in.As a result of this ,those organizations can make a wise decision of delivering and publishing more relevant news that a user is interested in and developing and promoting those articles which have a great impact for any user to share or comment.This prediction helps journalists and editors to efficiently allocate resources and create better reading experience.

This can also be used in trend forecasting and advertisements by understanding the human interests of which type of category has interested a user to have maximum number of shares.

II. Related works

Over the past few years,Researchers have conducted various web mining and machine learning models regarding the web content analysis and have come up with various conclusions.Among those studies few of them has been conducted for predicting the popularity of the online content. The works include predicting exact number of retweets on a tweet[4], the views for a YouTube video[1,2], ranking news articles[3] etc. The process of predicting is of two types - After publication prediction and Before publication prediction.

After publication technique is most common. Higher prediction rate is expected as all the information about received attention makes the task easier.Before prediction is challenging and effective .It uses only the metadata about the article to predict. There are many researchers who have worked on after publication but few have worked on before publication techniques.

Existing studies frame predicting problems as the regression, classification or clustering. There are many studies that worked on more than a model and compared their accuracies. Some of the models used includes linear regression, binary or multinomial logistic regression, Support Vector Machine(SVM), Random Forest, KNN,Naives Bayes,Gradient Boosting Machine etc. Some papers have used Gradient Boosting Machine(GBM) for before publication prediction on the same data. Their findings suggest that gradient boosting machine is able to predict popularity with a decent prediction rate using only statistical features associated with original news articles without using the original content of news articles or after publication attention. It was found that their model had 1.8percent improvement than previous model. They defined the popularity of an article based on a decision threshold . This was the first time that gradient boosting machine with linear and logistic regression loss functions is applied to the

task of predicting the popularity of online news. Some papers ranked news articles by predicting user comments using the linear model on a logarithmic scale and constant scaling model in after publication setting. They outlined that popularity prediction methods are the good alternative for automatic online news ranking. They have broke the articles of each dataset into small subsets where each dataset contains all articles published during a certain period of time before a specific reference hour and rank each subset of articles based on the number of comments that articles received after the reference hour.They have analyzed the efficiency of popularity prediction methods in the context of automatic online new ranking by evaluating two fundamental characteristics of online content: the distribution of popularity and the lifetime of articles. Their results indicate that a linear log popularity prediction model is an effective solution to online news ranking, with a performance that can evenly match more customized learning to rank algorithms.

A. Summary And Results of Literature Survey

Upon survey, one can come to a conclusion that Gradient Boosting Machine and Random forests algorithms are the most powerful for prediction of popularity. Random Forests performed well because the model is inherently robust to over-fitting and lends itself well to the high-dimensional data. With only a little bit of hyper-parameter tuning, it quickly outperformed the other models. Articles that have been shared for more than 1400 (median value) was labeled as popular.And also on projecting the data into PCA space they observed that the PCA dimensions do not lend themselves well in discriminating popular vs unpopular news articles and doesnt really give a clear picture of which features to be dropped or so.This is an important observation that one should take into picture as when one is trying to model on this dataset,rather than going with PCA for feature selection one should go with some other alternate solution for feature selection like stepwise selection or dropping the features step wise and so on.

And also based on the from Principal Components Analysis,they came to a conclusion that a non-linear model will perform better.When Linear Regression was used to predict the number of shares for each article, a very high MSE (Mean Squared Error) have made them to model the problem as a classification task.And choosing all the features had a poor result on the classification models so they did recursive feature elimination to obtain the top 20 features and the results were significantly different. And also some papers have come to a conclusion that the keyword related features have a stronger importance, followed by LDA based features and shares of Mashable links

And also,a lot of works were done on various different models and different accuracies were obtained .Some papers future works aims at not just prediction but also support aspects like article creation,evaluation of the prediction model for more complex and more unbalanced popularity dataset.One future work include predicting the life of an article.

But there was no single paper that focused on the recommendations that one can do to a particular article so that it can become popular before publication.Almost many papers were based on selecting different models and comparing which models gave good results in predicting the number of shares and the popularity of an article.But the point one should note over here is,there is no room left now in the selection of the model because almost all models gave the same accuracy rate.The only difference that changes the accuracy of the model is the way the feature selection is done.

III. Problem statement

1.In this work ,We will be using different models like Linear regression,logistic regression,SVM(support vector machines) model,Random Forest Model ,Gradient Boosting model,Decision trees,for estimating the number of shares for a particular article and comparing the accuracy of various models.

2.And also by checking the corrleation of various attributes with the number of shares,we will be answering questions like i)does having media elements (like images,videos) in the article affect popularity? ii)When does a article receive more shares.weekdays or weekends?Based on this one can suggest of when to publish a article. iii)which category of articles has maximum number of shares. iv)Predicting or finding the ideal value for the number of words in a title,content,keywords used,rate of positive words, rate of negative words and so on which makes an article more popular by using various visualization techniques like histogram (eg:-number of words in title vs number of shares) and finding which interval of values would really make a difference in the popularity.

3.One of the important work that we will be addressing in this project is recommending what changes does one need to make in one's article before publishing so that one can increase the popularity of that article.

A. Data Source

The dataset that we are using is provided by UCI Machine learning repository.This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years.It has total of 61 attributes among which 58 are predictive,2 are non predictive and a target attribute and 39644 instances.The nominal features like the which day of the week were been transformed with the common one hot encoding.Three types of keywords such as worst,average and best were captured by ranking all articles keyword average shares .The dataset also has a bunch of NLP features such as closeness to top latent Dirichlet allocation (LDA),title subjectivity,rate of positive and negative words and title sentiment polarity by using LDA to measure the closeness of current article to the previously computed topics.

IV. Explanation Of Each Component

A. Data Preprocessing

Usually data sets are highly susceptible to missing ,inconsistent values.Such low quality data might lead to bad results, hence we perform data cleaning to ensure that there is no

missing values.After this test was performed it was found that the dataset has no missing values.

We removed the unpredictive features from the dataset like the url,timedelta. We also removed the feature isweekend as it was a redundant feature.There was outlier in our dataset which we removed after going through the summary statistics and plotting boxplots.

The target variable in the given dataset is number of shares (a continuous variable).So we have introduced a new column into the dataset called Popular which has value either (0-not popular,1-popular).We considered an article to be popular if the number of shares is greater than or equal to the 75 percentage quantile value.

We performed log transformation on the target variable for doing linear regression for better results. And it seemed to give us improved results. As the log transformation is done for classification problem, we can say this problem turns out to be a good classifaction problem.

B. Data Visualization

Visualizations help to bring out a clear picture of the dataset.We did some visualizations to get some insights about our data. We plotted boxplots to check for outliers, histograms to reflect the relations between the articles becoming popular and title length,content length and others. We also plotted a histogram to see how does sharing of an article matter with respect to weekdays.Also, we could infer that a lot many articles were published during weekdays and not weekends.

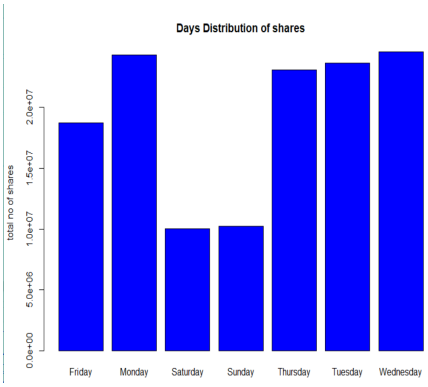

Fig. 1. day of the week vs total no of shares

C. Dimensionality Reduction

Reducing the dimension of the feature space is called dimensionality reduction.In terms of performance, having data of high dimensionality is problematic because (a) it can mean high computational cost to perform learning and inference and (b) it often leads to overfitting. Dimensionality reduction addresses both of these problems, while (hopefully) preserving most of the relevant information in the data needed to learn accurate, predictive models.There are various ways of doing Dimensionality Reduction.We in our project did:

1) Principal Component Analysis(PCA): It is a method that brings together a measure of how each variable is associated with one another(Covariance matrix) and the directions in which our data are dispersed(Eigenvectors) and the relative importance of these different directions(Eigenvalues).PCA combines the predictors and allows to drop the eigenvectors that are relatively unimportant.PCA is sensitive to outliers so before performing PCA we removed the outliers.Since PCA assumes the variables are linearly related but most of the variables are not linearly dependent in our dataset because of which the PCA couldn't give efficient results in reducing the feature space.

2) Step Wise Feature Selection: We randomly select a subset of features and train a model using them.Based on the inferences that we draw from the previous model, we decide to add or remove features from the subset.The problem is essentially reduced to a search problem and is computationally expensive.

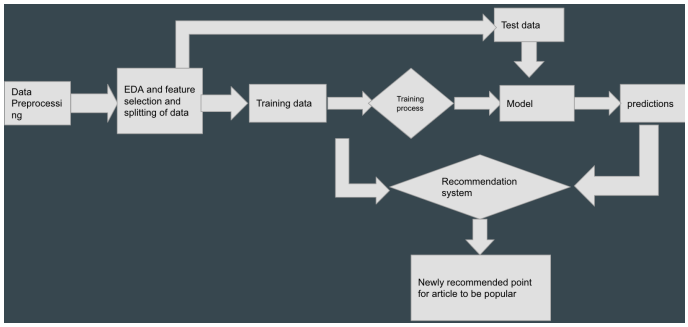

Fig. 2. flowchart

D. Build Machine Learning Models For Prediction

A prediction is often, but not always, based upon experience or knowledge.Prediction can be useful to assist in making plans about possible developments. Some good predictions saves lives and also reduce damage and economic loss. Prediction is useful in weather forecasting,election results prediction and so on. There are many predictive models that helps us in prediction. They use data mining and probability to forecast outcomes.Each model is made up of a number of predictor, which are variables that are likely to influence future works. In our project we have used models like linear regression,logistics regression,decision trees,random forest, gradient boosting method and support vector machines.

1) Linear Regression: It is a simple approach for prediction.It uses linear approach for modelling the relationship between a scalar response and one or more explanatory variables.In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.We predict the number of shares of an article using this model and then classify the article using the predicted value as popular if it is greater than the third quantile. However this model is not suitable. It requires independent variable and the dependent variable to a share linear relationship.

2) Logistic regression: The logistic model (or logit model) is a widely used statistical model that,in its basic form, uses a logistic function to model a binary dependent variable.It is a form of binomial regression. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Here we append a new variable to the dataset called popular whose value depends on number of shares.If the number of shares is greater than the 3rd quantile popular is set to 1 signifying the article is popular otherwise 0.In this model we try to predict the variable popular using binomial classification.This model results in probability of the popular variable to be 1 .The probability above 0.5 is considered to be popular(i.e, 1) otherwise not.The main challenge of logistic regression is that it is difficult to correctly interpret the results.

3) Decision trees: It is a type of supervised learning algorithm that can be used in both regression and classification problems.It works for both categorical and continuous input and output variables.It can capture the division better when compared to the logistic regression.We have used tree library and tree function to train the model.This model is prone to overfitting,especially when the tree is particularly deep and error due to bias and variance.

4) Random Forest: They are an ensemble learning method for classification,regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual tree. Random decision forest correct for decision trees' habit of overfitting to their training set.It reduces variance by training on different samples of the data. A second way is by using a random subset of features.To implement this model, we use randomForest library. The number of trees is assigned to be 100.

5) Gradient Boosting method: GBM is an ensemble learning algorithm that is the combination of gradient-based optimization and boosting. GBM produces strong prediction model by combining multiple weak prediction models (typically decision trees) in which weak models are created by sequentially applying to the incrementally changed dataset. We use gaussian distribution and the value of shrinkage is 0.01 and number of trees is 100. It predicts the values for different specified trees and find mean squared error.

6) Support Vector Machine: They are the supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Handwritten characters can be recognized using SVM.It can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.In our prediction we have used linear, polynomial and radial kernel functions. Radial kernel function has localized and finite response along the entire x-axis.

E. Recommendation System

There was no single paper or work that focused on the recommendations or the changes that one can do to a particular article so that it can become popular before publication as almost all papers were based on selecting different models and comparing the accuracies.

So we have come up with a recommendation system in this project apart from the models that were built for predicting the popularity. This particular recommendation system takes in the trained data which are popular,the test data point which is not popular and the K nearest neighbours that needs to be looked into .

So in this recommendation system , we calculate the euclidean distance of every trained data point with the test data point and we select the K nearest neighbours for that test data point. Once we get the k nearest neighbours ,we calculate the centroid of those k nearest neighbours and suggest the changes or recommend that centroid point as the changes that need to be made in the present test data point so that it can gain more number of shares.

But the main challenge or the disadvantage of this recommendation method is that in order to find the K nearest neighbours, one has to calculate the distance between every data points belonging to the trained data of popular category and the new data point and then selecting K smaller distance values.But if these number of data points are very huge ,then one has to calculate the distance for all those data points and then calculate the centroid.As a result,this operation is costly and time consuming.

V. RESULTS

After preprocessing the data, we performed exploratory data analysis and did some visualizations. After all these steps we modelled our dataset using the different machine learning models like linear regression,logistic regression,decision trees,random forest,SVM and GBM.

We started with linear regression on the target variable which gave us a large RMSE error. So, we performed log transformation on the feature and stepwise feature selection was done. This gave us a decent RMSE error and the accuracy of the model was also good. After that we moved to logistic regression which classified articles that are popular as 1 and unpopular as 0. This gave a better performance than the linear model. It gave 76% accuracy. Then we moved on to decision trees which made a tree to classify the articles based on other features. It gave a better accuracy than linear model but not logistic one. Further we went on to use Random Forest algorithm which built 500 trees to build our model. It turned out to be a good choice as it gave us an accuracy of 76.3% , better than other models discussed earlier. We also used GBM to model our data. It didn't give us a very satisfactory result. It did give very less RMSE. We used SVM also for building a model. We used 3 of its kernels-linear,poynomial,radial. This turned out to predict with atmost accuracy for our posed problem. SVM turned out to be the best model amongst all

discussed above for our dataset. Random Forest also faired well.

Models	Accuracy
Linear Regression	75%
Logistic Regression	76%
Decision Trees	75.9%
Random Forest	76.3%
Support Vector Machines(linear)	76.49%
Support Vector Machines(polynomial)	76.52%
Support Vector Machines(radial)	76.9%

Fig. 3. Models Comparision

We then picked some random unpopular data points from our dataset and used our algorithm to recommend some changes that can be made to the data point to become popular using k nearest popular neighbours.

Category	Original #Shares	#Shares after Recommending Changes
World	1300	18050
Technology	1000	7300
Business	777	178500
Entertainment	2200	4700

Fig. 4. Recommendation results

We also understood from the visualizations that the content length keeps on decreasing on increasing the content length in the news article and the optimal range was 0-500 words. Also, the title length optimal range was 9-11 words.

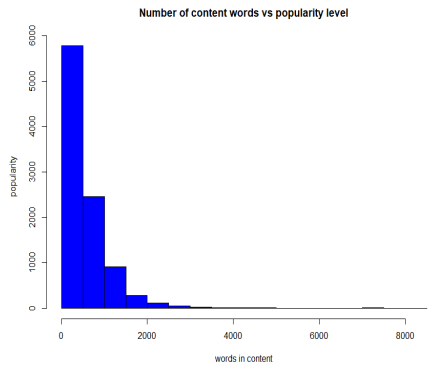

Fig. 5. title length vs popularity

VI. CONCLUSION

For our problem SVM turned out to be the best fitting model and gave better accuracy than others. The visualizations

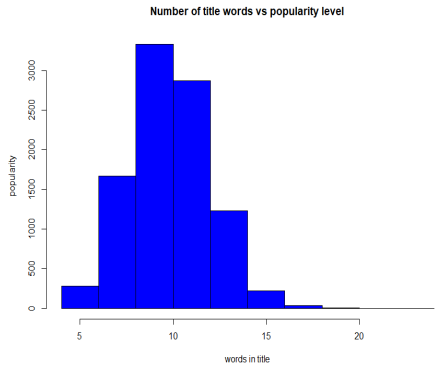

Fig. 6. title length vs popularity

helped us to infer that news articles with more content length won't fair well in the market. It should be within 0-500 words for optimal popularization. The articles are shared more on weekdays than on weekends. The title of the news article should not be very short or very long. Also, articles with media elements does contribute more to popularity of the article. Our recommendation algorithm works well in recommending the changes to be made to the unpopular article to become popular.

VII. CONTRIBUTION

Alka Simon -Contribution in preprocessing,visualizations,3 model implementations(Random Forest,Linear Regression,SVM) and Report making(Data Visualization,Results Analysis,Conclusion).

K Vennela-Contribution in building the recommendation system,Literature Survey and Report (Domain Introduction,Problem Statement,Recommendation System)

Yeshaswini M- Contribution in preprocessing,implementations of Logistic,Decision trees ,GBM models Literature Survey and report(Explanation of the components)

REFERENCES

[1] Szabo, G., and Huberman, B. A. Predicting the popularity of online content. Communications of the ACM 53(8):8088, 2010

[2] Gursun, G.; Crovella, M.; and Matta, I. Describing and forecasting video access patterns. In Proceedings of IEEE INFOCOM11, 1620, 2011.

[3] Tatar, A., Antoniadis, P., De Amorim, M. D., and Fdida, S. Ranking news articles based on popularity prediction. In Proceedings of ASONAM12, 106110, 2012

[4] A. Tatar, M. de Amorim, S. Fdida, and P. Antoniadis, A survey on predicting the popularity of web content, Journal of Internet Services and Applications, vol. 5, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1186/s13174-014-0008-y

[5] A. Tatar, P. Antoniadis, M. Amorim, and S. Fdida, From popularity prediction to ranking online news, Social Network Analysis and Mining, vol. 4, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1007/s13278-014-0174-8

[6] Md Uddin, Tanveer Ahsan et al Predicting the Popularity of Online News using Gradient Boosting Machine,University of Chittagong

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

[8] Daniel Reznikov, Sudeep Nagabhirava, Jaskaran Virdi, Shreyas Udupa Balckudru News Article Popularity - A Predictive Task