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*—We are trying to predict the popularity of an article(by estimating the number of shares a particular news article can get) based on the relevant features ,that we could extract from the dataset provided by Mashable. After predicting this,we will be recommending what changes can one 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. Introduction to Dataset

A. 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. The goal here is to predict the number of shares in social networking site.It has total of 61 attributes among which 58 are predictive , 2 are non predictive and a target attribute and 39644 instances.

III. 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.

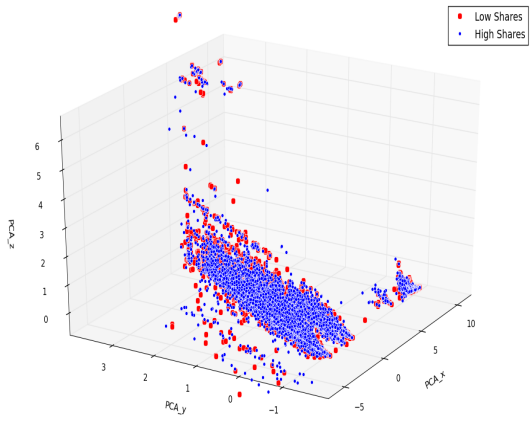
Existing studies frame predicting problems as the regression , classification or clustering. There are many studies that worked on more than a model and compared their accuracies. Some of the models used includes linear regression, binary or multinomial logistic regression, Support Vector Machine(SVM), Random Forest, KNN,Naives Bayes,Gradient Boosting Machine etc.

A. News Article Popularity - A Predictive Task by Daniel Reznikov and others at UCSD

1)Approach-[9]In this paper,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 step-wise selection or dropping the features step wise and so on.They have used three metrics to compare the models i.e Accuracy,area under ROC curve(for each of the 10 folds cross validation) and F1-score.

2)Summary of Results-By training the classifier on all predictive features, they obtained a mean AUC of 0.62, an accuracy of 53.07percent and an F1-Score of 0.24.

Fig. 1. PCA on 3 dimensions



3)Conclusion-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.

B. Predicting the Popularity of Online News using Gradient Boosting Machine by Md.Taufeeq Uddin and others

1)Approach-[7] have used Gradient Boosting Machine(GBM) for before publication prediction on the same data..GBM is an ensemble learning algorithm that is the combination of gradient-based optimization and boosting.

2)Summary of Results-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 there model had 1.8percent improvement than previous model. They defined the popularity of an article based on a decision threshold . They have used the logarithm of the original number of shares of news as the prediction label to train and test using GBM and RF regressors.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.

C. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News by Kelwin Fernandes and others

1)Approach-[8]The popularity of a candidate article is first estimated using a prediction module and then an optimization module suggests changes in the article content and structure, in order to maximize its expected popularity.They use a local search method for optimizing the article.This local search method is like the hill climbing,which iteratively searches within the neighborhood of the current solution and updates such solution when a better one is found, until a local optimum is reached or the method is stopped. And the search is only performed over a subset of features that are more suitable to be changed by the author and in each iteration, the neighborhood search space assumes small perturbations (increase or decrease) in the feature original values and suggests those changes.

2)Summary of Results-They adopted a rolling windows scheme with a training window size of W = 10,000 and performing L = 1,000 predictions at each iteration.And each classification model was trained 29 times (iterations), producing 29 prediction sets (each of size L).The random forest model was considered to be the best model among various models they used with an accuracy of 0.67, recall of 0.71 ,F1 of 0.63.

3)Conclusion-And also they have come to a conclusion that the keyword related features have a stronger importance, followed by LDA based features and shares of Mashable links.

D. A survey on predicting the popularity of web content by A. Tatar and others

1)Approach-Tatar et al[5] have described the different popularity prediction models, present the features that have shown good predictive capabilities, and reveal factors known to influence web content popularity.

2)Summary of Results-They summarise that most of the previous studies address the problem of predicting exact number if attention received by an article as future work.In addition to early popularity measures, several researchers studied the predictive power.They state that this direction is fully not explored and further works include powerful predictive features. For a web content with short lifespan, timely prediction is a real

Fig. 2. ROC curves for paper B

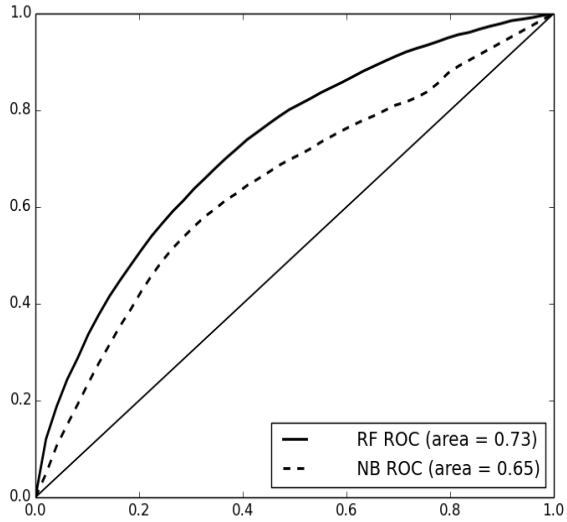
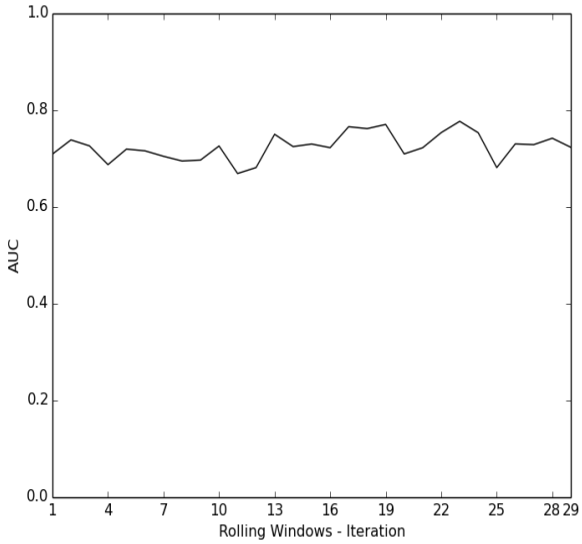


Fig. 3. AUC metric distribution over time for RF for paper B



challenge. A way to improve this is to extract recurrent events overtime and observe the level of interest that they generate and predict when these future events takes place.

E. From popularity prediction to ranking online news by A. Tatar and others

1)Approach- Tatar et al. [6] 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.

2)Summary of Results-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.Kaggle had also conducted competition for improvement of prediction method on the same dataset.

F. Complete Conclusion

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.And also PCA doesnt have much impact on feature selection so going with recursive feature selection (either step-wise feature selection or dropping the features) have a good result.And also ,a lot of works were done on various different models and different accuracies were obtained but no paper focused on the recommendations that one can do to a particular article so that it can become popular before publication.

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.

To do this there is very little room in selection of model but much in feature selection.It is done by considering every

single word in the article as an added feature.

IV. Proposed Problem statement

1.We will be using different models like Linear regression,Multi-classification model,KNN model,Random Forest Model for estimating the number of shares for a particular article and comparing the accuracy of various models.

2.checking the corrleation of various attributes with the number of shares,so that one can answer questions like

i)does having media elements (like images,videos) in the article affect popularity?If so, what should be the ideal number of media elements.

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 .One way of we thought of doing this is by using various visualization techniques like histogram (eg:-words in title vs shares) and finding which interval of values would really make a difference in the popularity.

3.So by estimating these ideal values for various attributes one can suggest of what changes does one need to make in one's article before publishing to make it popular.

V. Approach to be used

As a primary step ,we perform some EDA to gain an insight of the given data set. Having a first look of the data set we can infer that the dimensionality of the data set is quite large and choosing an appropriate attribute for the predictive model might be tough. As a preliminary step we perform Principal Component Analysis to identify variables in a data set that represent the most information about the data set but that does'nt give much information.

So we thought of pre-processing the data by step-wise feature selection method.

Later after doing the prediction of number of shares for an article,we would suggest what changes can one make in the features of an article in order for the article to become popular before publication. One of the solution that we thought of using is clustering the data points into different clusters or categories like popular ,not popular and so on based on the labels of the article.After clustering them,we would see K nearest neighbours of popular category from the new data point that we want to recommend the changes for.After knowing the K nearest neighbours ,we would calculate the "centroid" formed by those K nearest neighbours and recommend the centroid value as the new changes that one can do in one's news article features before publishing so that the article can obtain more number of shares.

A. Problems/Challenges one may face by this solution is:

- What K to select?
- In order to find the K nearest neighbours, one has to calculate the distance between the data points belonging to popular category and the new data point and then selecting K smaller distance values.But if the 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.
- Apart from this ,we would also answer to the 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?

  iii)which category of articles has maximum number of shares?

  iv)Finding the ideal value for the number of words in a title,content,keywords been used,rate of positive words to be used which makes an article more popular.

  One way of we thought doing this is by using various visualization techniques like histogram (eg:-words in title vs shares) and then coming to a conclusion of which interval of values would really make a difference in the popularity.

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] Zaman, T.; Fox, E. B.; Bradlow, E. T.; et al. A bayesian approach for predicting the popularity of tweets. The Annals of Applied Statistics 8(3):15831611, 2014

[5] 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

[6] 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

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

[8] 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

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