Predicting Popularity of Online News Reports

Nirmit Gupta(19),Bhavya Dubey(49),Meet Chaudhary(44),Harman Singh(50)

ABSTRACT:

With the help of Internet, the web news are often instantly spread around the world. Most of the people now have the habit of reading and sharing news online, for example , using social media like Twitter and Facebook. Typically, the news popularity are often indicated by the amount of reads, likes or shares. For the web news stake holders, it's very valuable if the recognition of the news articles are often accurately predicted before the publication. Thus, it's interesting and meaningful to use the machine learning techniques to predict the recognition of online news articles.

In our project, the dataset including 39,643 news articles from website Mashable, we attempt to find the simplest classification learning algorithm to accurately predict if a news story will become popular or not before publication.

Our research reflects that predicting news popularity with the help of tweets in social media platforms based on their shares and hashtags used during weekdays and weekends being done using different classification algorithms i.e linear regression, random forest, adaboost gives different result and the most accurate result obtained by our research is from Random forest classification algorithm giving an accuracy rate of 0.6743 which is highest amongst all the other classification algorithms we've used.

2) INTRODUCTION:

Working with machine learning algorithms within the large dataset is extremely common and particularly with the expansion of online news, it became very useful. Random Forest, linear regression and Adaboost are the common machine learning algorithms used for classification. during this research, we aimed to seek out the simplest model and set of features to predict the recognition of online news, using machine-learning techniques and implement various machine learning algorithms on the chosen features.The data source was Mashable, a well known online news website. Precision, AUC (Area Under the Curve) and F-measure were used to evaluate the results and their results were compared to seek out the foremost accurate amongst all. Random Forest seems to be the simplest model for prediction, and may achieve an accuracy of around 67%. Our work can help online news companies to predict news popularity before publication.

Various works have been done for prediction of online news popularity. In [1], the recognition of online articles is analyzed supported the user's comments. [2] defines the recognition in terms of a contest where the favored articles are those which

were the foremost appealing thereon particular day. Ranking Support Vector Machine (SVM) is employed to classify the appealing/non appealing of online news story . In [3], the amount of retweets is predicted using both the features of the retweet content (length, words, number of hashtag, etc.) and therefore the features of author (number of followers, friends, etc.). [4] collects a dataset with almost 40,000 articles from the Mashable website, compares five different methods on classifying popular/unpopular news articles and concludes that the Random Forest (RF) are able to do the simplest performance.

2.1) Data Exploration -

The dataset consists of 39,643 news articles from an online news website called Mashable collected over 2 years from the time period of Jan. 2013 to Jan. 2015. It is downloaded from UCI Machine Learning Repository as https://archive.ics.uci.edu/ml/datasets/Online+News+ Popularity# and this dataset is generously denoted by the author of [4]. For each instance of the dataset, it has 61 attributes which includes 1 target attribute (number of shares), 2 non-predictive features (URL of the article and days between the article publication and the dataset acquisition) and 58 predictive features as shown in Fig. 1. For example, the categorical features like the published day of the week and article category have been transformed by one-hot encoding scheme, and the skewed feature like number of words in the article has been log transformed.

Matched Source

Similarity 7%

Title: Online News Popularity Prediction | IEEE Conference ...

by F Namous · 2018 · Cited by 1 — Random Forest and Neural Network turn out to be the best model for prediction, and both of them can achieve an accuracy of 65% with optimal parameters. Our work can help online news companies to predict news popularity before publication.

https://ieeexplore.ieee.org/document/8649529