Predicting popularity of online news articles

Aditya Bugalya
IIIT-Delhi
aditya19457@iiitd.ac.in

Ashwin Tomer
IIIT-Delhi
ashwin20289@iiitd.ac.in

Rupin Oberoi
IIIT-Delhi
rupin20571@iiitd.ac.in

Sarthak Dixit
IIIT-Delhi
sarthak20574@iiitd.ac.in

Abstract

We aim to devise a way to accurately predict the popularity of news articles on the basis of various characteristics such as its textual and multimedia content. Identifying important features influencing the popularity of an article would help writers to tweak their content in order to make it more engaging. This is beneficial not only for such content creators but also for the consumers, and would also lend key insights into general human psychology. In this project we make use of both regression and classification algorithms to make predictions.

Link to Github repository

1. Introduction

With increasing access to cheaper internet, people's consumption of online sources for their information is at an all-time high. To capitalize on this growing demand, several news organizations now primarily focus on this channel of broadcasting. Hence it becomes essential for these organizations to differentiate from others in order to attract higher readership leading to higher revenues from advertising and subscriptions.

We aim to develop a reliable way to predict the popularity of a news article on the basis of the number of times it has been shared by the readers to their friends and family.

With the massive amount of such articles that are published everyday and only a few getting substantial attention from the general populace, it is hard to explain the reasons behind it. Hence, we use various machine learning techniques relating to regression and classification to predict the popularity of an article before its publication on the basis of various features. For this purpose, we use the dataset provided by UCI [1] which contains a wide variety of features from over 39000 articles published by Mashable.com, a prominent digital media platform.

2. Literature Review

Roja Bandari et al. in their paper 'The Pulse of News in Social Media: Forecasting Popularity' [2] have addressed the problem of predicting popularity of news items on social media prior to its release. They have analyzed the number of times an article's URL is mentioned or shared on Twitter. They consider characteristics such as the news source, article category, subjectivity of language used and named entities mentioned. They assign a score to each news source based on the number of articles published and t-density which is similar to the average number of shares per article. On calculating a similar metric on the basis of the reputation of the news sources it was found that there it has an insignificant correlation with the popularity.

Carlos Castillo et al. in their paper "Characterizing the Life Cycle of Online News Stories Using Social Media Reactions" [3], present the study of life cycle of a news article. They took a few uneventful weeks to study articles of Al Jazeera to avoid temporary peaks in views. There are two types of news articles, breaking news and in-depth, and the way people found these articles was classified into 4 types: internal links (website home page etc.), external links (shared publicly on social networking sites (SNS)), direct links (sent via messages or emails) and search results. To study the life cycle, several variables were computed such as number of Facebook shares/minute, number of tweets/minute, number of followers of the people who tweeted them etc. Upon studying the patterns in the life cycle three different types of graphs were most prevalent: Decreasing(78%), Steady or Increasing(12%) and Rebounding(10%). Decreasing views articles are mostly of the breaking news type for example about a sports match, although the views are high the shelf-life of the article is very low. Steady or Increasing articles are mostly in-depth articles that talk about long standing issues and are constantly being talked about in social media. Rebounding articles are those which were on a steady decline but saw a sudden spike in views due to internal (Main focus of home page/a related popular article was released) or external links (raised by an influencer on SNS). Looking at these trends and wanting the best of both worlds, reporters figure out how to bridge a breaking news to a long standing issue to ensure both views and long shelf life. Seeing which in-depth articles are gaining popularity also helps traditional TV news in figuring out what to show to keep up the viewership in case of no breaking news.

3. Dataset description

3.1. Attributes and visualization

The dataset contains information about articles published by Mashable, and has been provided by the UCI archive. It consists of 39,644 samples with 58 predictive features, 2 non-predictive features (URL and time_delta) and a target variable (shares).

Category	**Description**
General article characteristics	Article Category (technology, business etc); Day of publication
Textual content and NLP related	Number of links, tokens, average subjectivity, sentiment polarity, closeness to Latent Dirichlet Allocation (LDA) topic
Multimedia content	Number of images and videos
Titular content	Subjectivity and sentiment polarity of title
Non- predictive	Article's URL; time difference between publication and dataset acquisition
Target	Number of shares

Table 1: Broad categories of dataset attributes and description of corresponding features

The dataset consists of binary and numeric attributes and no categorical attributes, as all attributes which could be categorical such as day of the week, or category of article are already present in one-hot encoding form.

3.2 Preprocessing

3.2.1 Fundamental preprocessing

No null values were found in the dataset. Some basic preprocessing such as removing typographical errors, extra spaces from column names was done. We checked for invalid values such as negative, non-integer values in the appropriate columns and no such values were found.

Samples corresponding to articles not older than 2 weeks at the time of dataset acquisition were removed as there might be some readership and sharing of such articles. This brought down the number of samples to 39,431.

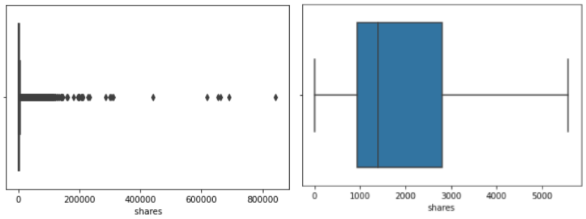


Figure 1: Boxplots depicting distribution of values for the number of shares before and after handling outliers

3.2.2 Handling outliers

As can be seen from the boxplot in Figure 1, there are a significant number of outliers for the number of shares and the magnitude is very high as well. This can be explained by the fact that there are some articles which gain huge popularity sometimes by getting shared by an influential person on a social media platform.

Outliers are defined as samples which lie outside the range [Q1 – (1.5*IQR), Q3+(1.5*IQR)], where Q3 is the value of the 75th percentile sample and Q1 is the value of the 25th percentile sample. IQR is the inter- quartile range i.e. Q3 – Q1. We had 2 options to handle these, either to ignore such samples or to cap the values at the upper bound for non-outliers. We chose the latter as the former would have led to loss of significant number of samples and information.

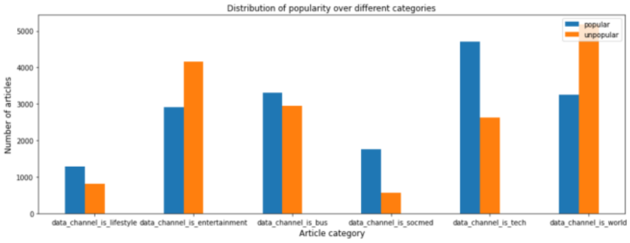


Figure 2: Proportion of popular news articles for every category (classified according to median being the deciding factor for popularity)

3.2.3 Converting URLs to article titles

We realized that although the URL column has been classified as a non-predictive feature, we could still derive some information from them as URLs contain a part of the article title. We converted the URL into a normal string similar to the title using regex and tokenized them using the NLTK library. This can be used for some NLP related tasks which we would like to explore in the future.

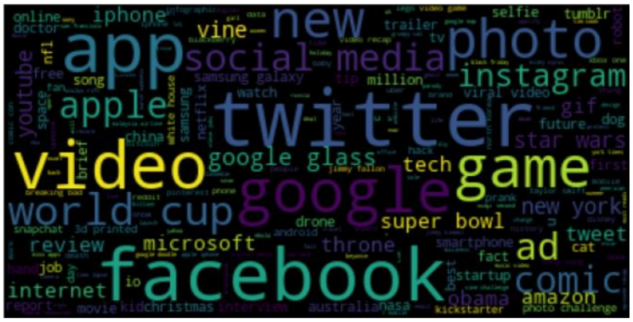


Figure 3: Word Cloud demonstrating the most common words in article titles weighted to logarithm of article's number of shares.

3.2.4 Feature selection

We plotted a heatmap of all predictive features with each other, and identified pairs of features which were highly correlated and removed one of them. This helped in eliminating 4 features: 'n_non_stop_unique_tokens', 'n_non_stop_words','kw_max_min', 'kw_max_max'.

4. Methodology

We have classified articles with greater than or equal to the median number of shares as popular and the rest as unpopular. Hence, the problem can now be approached both as a regression and a binary classification task.

We have split the dataset into training and testing set in the 80:20 ratio for evaluating various regression and classification algorithms used.

4.1 Regression

We have used linear regression for predicting the number of times an article will be shared using the preprocessed data. The results can be for the same can be found in Table 2.

From the results it can be inferred that there is high bias and low variance which implies underfitting. We also tried Lasso and Ridge regression but expectedly, because of underfitting, the results did not improve even after choosing the regularization parameter using Sklearn's GridSearchCV.

4.2 Classification

We have classified an article as popular if it has greater than the median number (1400 in this dataset) and rest as unpopular. We have used several classification algorithms as follows:

4.2.1 Naïve Bayes

Naïve Bayes classifier is one of the fastest and simplest classification algorithms, which works very efficiently for high dimensional datasets, like the one we are currently working on. We consider it as a baseline model for our classification problem.

4.2.2 Logistic Regression

Logistic Regression performs well and does not make any assumptions about distribution of classes in the feature space. Hence making it a great choice for our binary classification problem as the distribution is varied for different features.

4.2.3 Decision Tree and Random Forest

Decision trees are among the most interpretable classifiers, and also give decent results. By using grid search we found the best performing hyperparameters: max_depth = 6 and criterion: entropy

Random Forests classifier is based on the concept of the bagging algorithm, it is very stable and can handle outliers very efficiently and is not impacted by noise very much.

4.2.4 Support Vector Machine

Support Vector Machine (SVM) is a commonly used classifier which works by determining the optimal separating hyperplane between the support vectors from the samples of the two classes. Radial basis function (RBF) kernel gave the best results, as it transforms the input samples in a more complex and non-linear fashion leading to more separation of most data points causing the decision boundary to be a better separator. Using grid search, the optimal regularization parameter was 0.1 which implies the need for higher regularization.

4.2.5 Artificial Neural Networks

Artificial Neural Networks (ANNs) are the most basic form of deep learning architectures, in which the multiple layers of perceptrons are present and learning is through backpropagation of loss through these layers. Complex decision boundaries can be created by using more layers in a network, and hence ANNs are a powerful classifier. On trying various hidden layer configurations we found hidden layer sizes 40,32,20,12,8 as relatively better performing.

4.2.6 K-nearest neighbors

K-nearest neighbors is an unsupervised training algorithm as it does not require the labels of samples. KNN takes parameter K and creates K centroid points, and every sample is associated with one of these depending on the distance. After every iteration these centroids are updated until convergence which is certain to happen after a finite number of iterations.

5. Results and Analysis

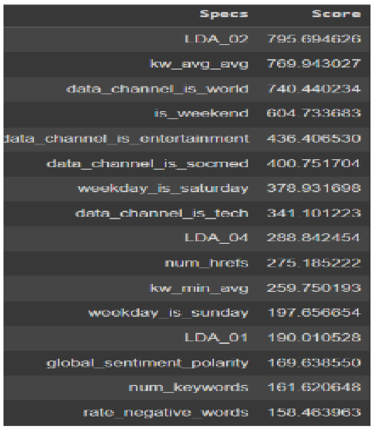
The high bias and low variance in linear regression clearly indicate underfitting, this is also confirmed by the lack of significant change when regularization is used. This can be due to the significant variance still present even after capping the outliers.

The results for Naïve Bayes classifier are expectedly the least desirable and thus serves as a baseline classifier for our task. The metrics are the best for Random Forest (n_estimators = 25, criterion = entropy) for which the hyperparameters were chosen using grid search. However due to the extremely high accuracy on the training set, there is overfitting happening which was to reduced to some extent by being more in feature selection using their importance. We determined most important features using Sklearn’s SelectKBest function and features corresponding to day of the week, article category, global (all over the article) sentiment, number of negative words used were among the ones with the highest scores.

A high number of iterations were required for convergence which was thought to be because of the non-scaling of the data, however scaling did not help much as it led to worse results.

The results for training and test set are mostly identical implying that there is no overfitting, which can be the case when SVM is applied on high dimensional data.

The results with KNN, ANN and SVM were not significantly better than Naïve-Bayes in terms of accuracy which might mean that as these algorithms focus on the distinguishability of the data points as a whole and not on particular attributes such as decision trees (which gave best results), the data points are mostly similar with only some features being discriminative.



Specs	Score
LDA_02	795.694626
kw_avg_avg	769.943027
data_channel_is_world	740.440234
is_weekend	604.733683
lata_channel_is_entertainment	436.406530
data_channel_is_socmed	400.751704
weekday_is_saturday	378.931698
data_channel_is_tech	341.101223
LDA_04	288.842454
num_hrefs	275.185222
kw_min_avg	259.750193
weekday_is_sunday	197.656654
LDA_01	190.010528
global_sentiment_polarity	169.638550
num_keywords	161.620648
rate_negative_words	158.463963

Figure 4: Most important attributes with scores

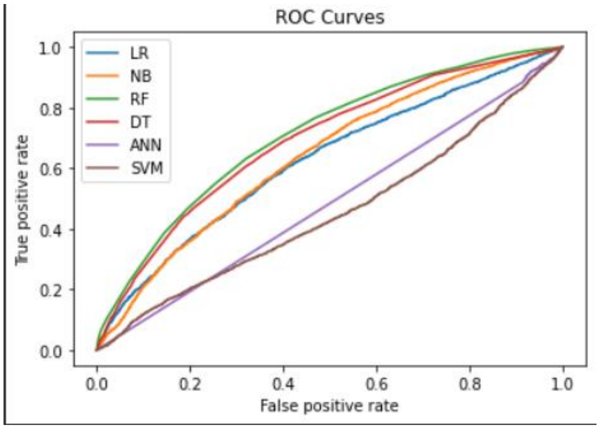


Figure 5: The ROC curve for the classifiers has been demonstrated, with Naïve Bayes being almost as good as random selection.

	R^2	RMSE
Linear regression		
Training set	0.125	1521
Testing set	0.12	1513
Lasso regression		
Training set	0.119	1531
Testing set	0.114	1528
Ridge regression		
Training set	0.12	1529
Testing set	0.113	1526

Table 2: Results of various methods used for regression

Model	Accuracy	Precision	Recall	F1-score
Test set				
Naïve Bayes	0.51	0.16	0.71	0.26
Logistic Regression	0.61	0.67	0.62	0.65
Decision Tree	0.63	0.69	0.64	0.67
Random Forest	0.66	0.70	0.66	0.68
Support Vector Machine	0.53	0.99	0.53	0.69
Artificial Neural Network	0.52	0.95	0.53	0.68
K nearest neighbors	0.53	0.34	0.6	0.43
Training set				
Naïve Bayes	0.51	0.15	0.68	0.25
Logistic Regression	0.61	0.68	0.62	0.65
Decision Tree	0.64	0.70	0.65	0.68
Random Forest	0.99	0.99	0.99	0.99
Support Vector Machine	0.53 (0.57 with high number of iterations)	0.999	0.53	0.69
Artificial Neural Network	0.53	0.95	0.53	0.68
K nearest neighbors	0.76	0.56	1	0.72

Table 3: Results from various classifiers

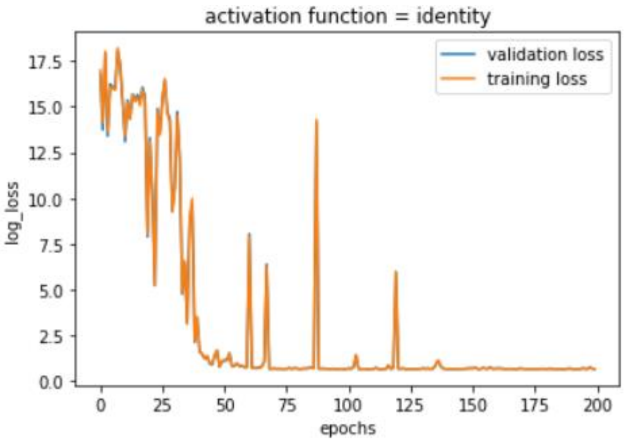


Figure 6: Training and validation losses for ANN v/s epochs

6. Conclusion

6.1 Learnings

- Exploring ongoing research about the task at hand, and identifying results relevant for us
- Preprocessing on a large dataset
- Handling outliers
- Analyzing various models on the basis of various metrics
- Working as a team

6.2 Individual contributions

Aditya did the EDA and trained various models, while Ashwin was responsible for literature review and in analyzing the results from the models trained. Rupin also did literature review along with data preprocessing and making the report, and Sarthak did feature selection and cleaning the data.

6.3 Future opportunities

We could not get access to the text of the articles, with the advent of transformers, a new paradigm in deep learning architecture, which could help in better results as they are capable of capturing context over large texts such as news articles. Another direction to look for will be the thumbnails used in the article and extract some features such as the variance and types of colors used.

7. References

[1] 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 Artificial Intelligence, September, Coimbra, Portugal.

[2] Bandari, R., Asur, S. and Huberman, B. 2021. The Pulse of News in Social Media: Forecasting Popularity. *Proceedings of the International AAAI Conference on Web and Social Media*. 6, 1 (Aug. 2021), 26-33.

[3] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. Characterizing the life cycle of online news stories using social media reactions. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). Association for Computing Machinery, New York, NY, USA, 211–223. https://doi.org/10.1145/2531602.2531623