# Online News Popularity Prediction

Abhijit Kannepalli, Bhavesh Chatnani, Hemil Kothari, Saurabh Saoji, Shrey Shah

School of Information Sciences, University of Illinois at Urbana-Champaign

IS 507: Data, Statistical Models, and Information

December 10, 2022

# Online News Popularity Prediction


## Abstract

One does not need a fortune teller to predict if the news is going viral; all you need is this ML model. The maximum growth in the attention given to a particular news piece is referred to as the article's popularity. The various types of news can be read on social networking sites and websites. The popularity of online news is determined by various elements, including the number of shares on social media, the number of comments by website users, and the number of likes.

However, in this case, we determine the popularity using the number of shares by setting a threshold limit. As a result, developing an automated decision support system to forecast the level of interest in news is essential, given that this will aid in gathering business intelligence. By utilizing several machine learning algorithms, the work that will be provided in this research aims to predict the number of news article shares above or below the threshold.

At first, PCA is utilized in order to scale down the dimensions. In the following step, a K-means Clustering, and XGBoost algorithm, are put into action to categorize the news sentiments and forecast the level of interest in the news. The system's functionality is evaluated based on the dataset obtained from the machine learning repository at UCI. The effectiveness of each of the three approaches in prediction is investigated by considering assessment metrics. XGBoost gives an accuracy of 71%, considering a limit above 2000 shares as popular.


## Introduction and Dataset

With the advent of the internet and social media, online news has become integral to everyone’s life. The most significant increase in attention paid to a specific news piece is known

as news popularity. For reading various news stories, people use social networking and news websites. It is a medium through which everyone keeps themselves updated about the recent happenings worldwide. A lot of variables, including the quantity of social media shares, the number of visitor comments, the quantity of likes, etc., influence the popularity of online news. Building an automated decision support system to forecast news popularity is therefore vital because it will aid in business intelligence. An area that we are typically trying to focus on is the factors influencing the sharing of news articles.

The dataset that we are using is from the UCI Machine Learning Repository, which contains the data of 39797 news articles published by Mashable over a period of two years. There are 61 features, including the number of shares for each article. It contains details such as the number of images, videos, and keywords in an article. Positive and negative polarity ratings, the day the article was published, and the channel through which the article was published are also included. Through this research, we are trying to predict the number of times an article might get shared and categorize it as either a popular or a common article, segregate articles based on their polarity, and identify relationships of factors with the number of shares.

## Literature Review

(Omar, 2007) seeks to observe how the transition from conventional news to online news has occurred and which factors have contributed to online news becoming popular. Immediacy is a factor that can be considered one which has been shown to influence both instant gratifications for the reader and allow readers to have faster access to the latest issues. This also means a certain news category will spread faster than others, which can be determined by understanding the polarity. (Shirsat et al., 2017) successfully estimates the polarity of words (positive, negative,

neutral). Using text mining and Term Document Matrix, sentiment analysis is performed at document and sentence levels so that if the sentence is positive, the document is positive, and if the document is negative, the sentence is negative. The articles were divided into genre groups based on their sentiment scores and classified as positive, negative, or neutral. Unlike the previous two journal articles (Deshpande, 2017) identifies the optimal model to predict web news popularity by building an automated decision support system without considering the effect of sentiments or polarity of the article content. (Omar, 2007) helps us understand the hypothesis testing that could be used. The application of the polarity of words and its analysis can be determined by (Shirsat et al., 2017), and the various machine learning methods that can be used to predict the accuracy of our model can be determined by (Deshpande, 2017) which can further be improved.

Firstly, before performing any research techniques on the data, we intend to find out the most critical features that can be used to perform the analysis. The study done by (Deshpande, 2017) related to the Linear Discriminant Analysis technique for performing dimensionality reduction can be useful here. Next, we will be working on predicting the number of shares/popularity of online news by using appropriate machine learning methods. A hypothesis that we will work on is whether the type of data channel (entertainment, lifestyle, etc.) influences the number of shares of the news articles. Another hypothesis would be to check if there is any association between when the online article is read and its number of shares.

We hypothesize that authors focus more on writing negative articles due to a common belief that negative articles grab more attention. We will work on categorizing these articles into various categories such as Positive, Negative and Neutral and test our hypothesis based on the size of obtained clusters. The distribution among data channels will be more or less uniform since we

do not have any information about the types of readers and their preferences. The number of unique words, the number of images, the length of the article, and the sentiment (both positive and negative) can be some of the key features in deciding the popularity of an article. Because people are more relaxed on weekends, an article published on a weekend will be more popular than one published on a weekday. We plan to learn the impact of negative and positive sentiments on news consumption and understand how it can affect the shareability of the news. This study can help us understand the type of content to be incorporated into online articles to ensure maximum outreach.

## Methods

Before performing any research techniques on the data, we intend to perform dimensionality reduction using Principal Component Analysis (PCA) and identify the reduced components. PCA is a suitable technique that can be retain the information present in the original dataset using a smaller set of features. We perform PCA with promax rotation using 5 components. We will predict the number of shares/popularity of online news using appropriate machine learning methods. XGBoost can be utilized to ascertain this, as this technique provides regularized learning that helps reduce the loss function, preventing overfitting. Boosting may assume an ordinal relationship between the encoded values for the input variables. Additionally, XGBoost can only handle numeric vectors. If any of these assumptions are found in the dataset, we shall encode the categorical variables to make them numerical. We hypothesize that we can predict the shares of online news accurately. We will find the amount of articles based on the polarity of the words in articles, which may be related to positive, negative, or neutral sentiments. This will be done by using the K-means clustering technique.

A hypothesis that we will work on is whether the type of data channel (entertainment, lifestyle) influences the number of shares of news articles. Combining the data channel categories, we can get a single variable whose attributes will be categorical. Since our outcome variable is continuous, we can use a one-way ANOVA or Kruskal-Wallis test to determine the validity of the hypothesis. The data will be checked for normality and run the test. Another hypothesis would be to check any association between when the online article is published and its number of shares. We can run unpaired T-Test or Mann-Whitney tests to extract a potential pattern by comparing the outcome variable with various binary outcomes that we have in our hands. The data will be checked for normality and variances, upon which we will determine the final test we will proceed with.

## Results and Discussion

For PCA, we have checked collinearity using a correlation plot of variables that signifies that there is sufficient correlation among the variables. We have run the Anderson-Darling test for checking the normality and skewness. For non-normal data we have applied log and square root transformations. We have performed listwise deletion to remove missing values. We have scaled the data to standardize the variables. The appropriateness of the data was determined using Bartlett's test and KMO measure of sampling adequacy. The KMO test revealed a value of 0.5, while Bartlett's test was 12926832 ($df = 44$, $p < 0.001$). The internal consistency among the features was determined using Cronbach's alpha coefficient which was 0.53. Two major criteria considered for factor retention were scree plot and eigenvalues greater than one. PCA with promax rotation was performed using 5 components. The 5 components obtained were interpreted based

on the input features and renamed as Number of Tokens, Self Reference Shares, Keyword based Shares. Influence of Positive News, and Non-Stop Words and Token Length (Table 1)

K-Means clustering was performed on the dataset. Factors such as average positive and negative polarity, minimum and maximum positive and negative polarities, and title subjectivity were taken into consideration. It was found that 78% percent of the articles had subjective titles(Figure 3) and polarizing content(positive or negative). Only 22% of the articles can be classified as neutral articles. Overlapping was found in the cluster plot (Figure 4). This tells us that even though an article might be overall negative but it also has some positive elements in it. A preliminary analysis was done before performing XGBoost, such as checking if the variables are numeric. Prediction using XGBoost resulted in an accuracy of 71% (Figure. The total number of false positive points was 350, and false negative ones were 1569, assuming the threshold of the number of shares to be 2000. The sensitivity (true positive rate) came to be around 0.72, and the specificity was lesser at 0.62, and kw_avg was the most significant variable. The data channels explain around 12% of the data.

The results from the Kruskal-Wallis test to determine if data channels and shares were dependent on the day of the week the article is published on the transformed dataset showed a p-value much less than 0.001, which tells us that these factors are not associated with the final number of shares.

A point of difference obtained from the above analyses is with respect to the importance of the type of channel a news belongs to. The XGBoost technique considers the data channel as an important feature in making the decision about the news popularity. On the other hand, the result of the Kruskal-Wallis test suggests that the data channel is not significant in determining the popularity of news based on its shares.

# Conclusion

The research focuses on improving the performance of news popularity prediction models. It is possible to measure the popularity of news by a variety of metrics, including the number of likes, comments, and shares. The 'number of shares' of online news stories on social networking websites is employed as the predictor variable for popularity in this thesis. People disseminate news that they believe should be made public. Analysis is performed on XGBoost strategy for predicting popularity, taking into account various situations and evaluation metrics. The accuracy improved to about 71%.

In this work, these values are enhanced, resulting in improved classification performance when compared to existing research methodologies. Machine learning techniques such as K-means clustering, XGBoost along with dimension reduction techniques like Principal Component Analysis, are explored and deployed on a dataset of news popularity to yield a high predictive model. PCA is utilized for dimension reduction. XGBoost outperforms other models based on past research. This study demonstrates that XGBoost is an efficient model for predicting the popularity of news articles. From the experimental results, it is clear that the suggested algorithm is a useful tool for addressing the problem of popularity prediction.

# References

Deshpande, D. (2017). Prediction & Evaluation of online news popularity using Machine Intelligence. *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. https://doi.org/10.1109/iccubea.2017.8463790

Omar, B. (2007). The Switch to Online Newspapers Could Immediacy Be a Factor?

Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2017). Document level sentiment analysis from news articles. *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. https://doi.org/10.1109/iccubea.2017.8463638

# Tables and Figures

**Figure 1**

*Scree Plot for finding the optimal number of components for PCA*

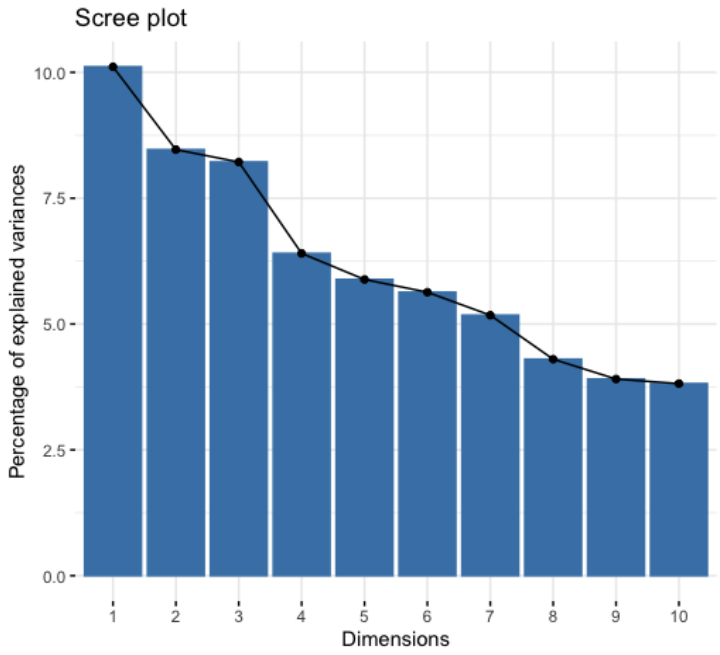


**Table 1**

*Components of Dataset after Principal Component Analysis*

| Item | Mean | Standard Deviation | Factor Loading | Eigen value | % Variance Explained |
|---|---|---|---|---|---|
| **Component 1: Number of Tokens** | | | | 4.55 | 15.44 |
| n_tokens_content | 21.62 | 8.26 | 0.87 | | |
| n_unique_tokens | 0.54 | 0.10 | -0.90 | | |
| n_non_stop_unique_tokens | 0.69 | 0.10 | -0.84 | | |
| num_hrefs | 2.94 | 1.24 | 0.64 | | |
| num_imgs | 0.02 | 0.76 | 0.68 | | |

| | | | | | |
|---|---|---|---|---|---|
| **Component 2: Self Reference Shares** | | | | 3.81 | 14.81 |
| self_reference_min_shares | 37.38 | 41.40 | 0.88 | | |
| self_reference_max_shares | 59.18 | 66.12 | 0.91 | | |
| self_reference_avg_sharess | 4868.44 | 20024.57 | 0.81 | | |
| self_reference_avg_shares | 5.94 | 3.64 | 0.89 | | |
| **Component 3: Keyword based Shares** | | | | 3.69 | 14.74 |
| kw_min_min | 2.47 | 1.81 | -0.73 | | |
| kw_min_max | 43.20 | 72.49 | 0.58 | | |
| kw_max_max | 626708.30 | 286117.20 | 0.83 | | |
| kw_avg_max | 393.30 | 146.59 | 0.89 | | |
| kw_min_avg | 847.12 | 1008.16 | 0.59 | | |
| kw_avg_avg | 51.01 | 8.72 | 0.9 | | |
| **Component 4: Influence of Positive News** | | | | 2.88 | 14.17 |
| global_sentiment_polarity | 0.13 | 0.08 | 0.88 | | |
| global_rate_positive_words | 0.20 | 0.03 | 0.66 | | |
| rate_positive_words | 0.73 | 0.15 | 0.93 | | |
| rate_negative_words | 0.50 | 0.13 | -0.94 | | |
| **Component 5: Non-Stop Words and Token Length** | | | | 2.65 | 10.46 |
| n_non_stop_words | 0.99 | 0.05 | 0.90 | | |

| | | | | |
|---|---|---|---|---|
| average_token_length | 4.64 | 0.39 | 0.89 | |

**Figure 2**

*Cluster means table for identifying the clusters*

```
K-means clustering with 3 clusters of sizes 5931, 3061, 4908

Cluster means:
  avg_positive_polarity min_positive_polarity max_positive_polarity avg_negative_polarity min_negative_polarity
1             0.1128972          -0.18064712             0.2645179            -0.3098292            -0.3491159
2            -0.5575088           0.43637679            -0.9504595             1.0635771             1.1072897
3             0.2112758          -0.05385723             0.2731257            -0.2889186            -0.2687057
  max_negative_polarity title_subjectivity title_sentiment_polarity abs_title_subjectivity abs_title_sentiment_polarity
1            -0.1275734         -0.7807579               -0.1796215             0.70382189                   -0.5083924
2             0.4567408         -0.2578490               -0.1668563             0.07895857                   -0.3748514
3            -0.1306939          1.1043095                0.3211251            -0.89976769                    0.8481450
```

**Figure 3**

*Graphical representation of Clusters*

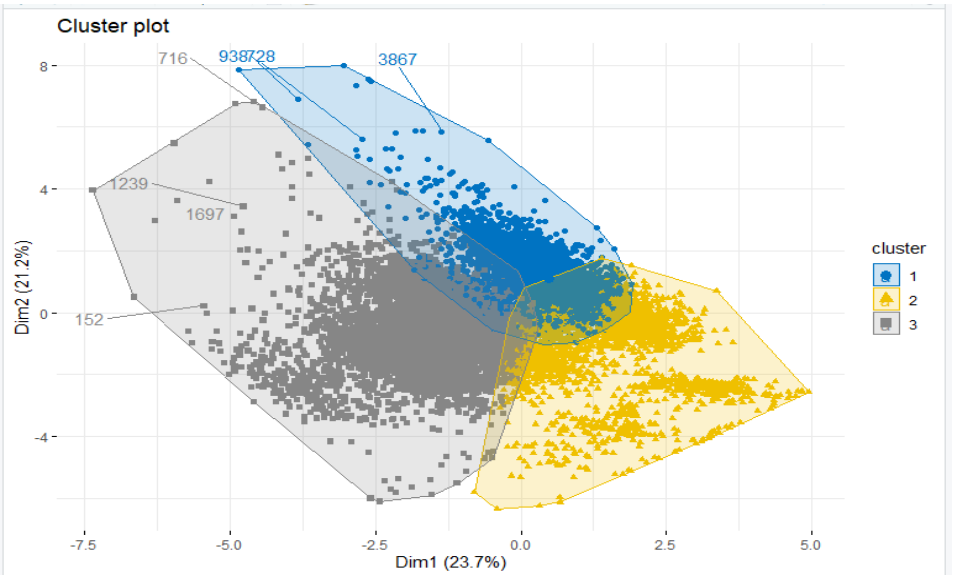

**Figure 5**

*Accuracy and Confusion Matrix*

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4197  350
         1 1569  586

               Accuracy : 0.7137
                 95% CI : (0.7027, 0.7245)
    No Information Rate : 0.8603
    P-Value [Acc > NIR] : 1

                  Kappa : 0.229

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7279
            Specificity : 0.6261
         Pos Pred Value : 0.9230
         Neg Pred Value : 0.2719
             Prevalence : 0.8603
         Detection Rate : 0.6262
   Detection Prevalence : 0.6785
      Balanced Accuracy : 0.6770

       'Positive' Class : 0
```

**Figure 6**

*Significant variables using XGBoost*