CAPSTONE PROJECT

INTERIM REPORT

on

PREDICTING POPULARITY OF ONLINE NEWS ARTICLE

Under the guidance of

Ms. Vidhya K

Submitted by

Akash P Bhatt

Swapnil Wagh

Saket P Shinde

Nicholas Lee D'Souza

Asmita Dileep Ghoderao

Table of Contents

Industry Review

The consumption of online news accelerates day by day due to the widespread adoption of smartphones and the rise of social networks. Dynamic news articles in today's times have a very short lifespan. In this volatile era, articles need to reach a large population of users for it to be popular. It is a known fact that articles which appeal to a broad section of users achieve popularity and become viral, although the popularity of the articles can either be short-lived or have a longer life span than expected.

News reporting and broadcasting online, have become a lucrative asset for news agencies due to a large number of users being exposed to news articles in real-time. This enables the news agencies to spend more time and resources on factors that influence the popularity of news articles. For example, a social media handle that serves as entry point for articles that have the potential to reach a large audience would be assessed by news agencies to understand the factors that drive popularity.

It becomes vital for news agencies to predict the popularity of online news articles by analysing its content, finding the factors that influence its popularity and ways it is consumed by users. This leads to the creation of more relevant, user-centric content by news agencies as well as effective allocation of resources to target and create news content.

In the recent years, there has been a lot of research done to predict the popularity of news articles published by Mashable. From the previous researches done, it is evident that ensemble models and non-linear classifiers are preferred for predicting the number of shares or popularity of news articles for the current dataset.

Furthermore, analysis of news content is also beneficial for trend forecasting, understanding collective human behaviour, enabling advertisers to propose more profitable techniques to target users, and finding the right users to consume the articles.

Literature Survey

Over the last couple of years, a lot of research has been conducted in order to predict the popularity of online news content. Two approaches of popularity prediction techniques suggested in the papers are:

- After Publication: A more common technique, which uses features capturing the attention that one content receives after its publication. Here the utilization of information about the received attention makes the prediction task easier.

- Before Publication: It is a relatively challenging and effective technique. This technique uses only content metadata features that are known prior to the publication of contents instead of using features leading to the attention that article receives after its contents are released. The prediction is more desirable as far as it fosters the possibility of decision making to customize the content before the release of content.

1. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News

   This paper explains the data collection process and the approach towards solving the problem by developing an intelligent decision support system which not only predicts the popularity of news articles but even optimizes a subset of features to increase the popularity. They classified attributes into: number, ratio, bool, nominal. The attributes related to number of keywords were log transformed and one-hot encoding was applied to nominal data. They also took into consideration the Natural Language Processing attributes such as LDA topics, text/title subjectivity and polarity. The regression problem was converted to a binary classification problem with classes as popular and unpopular. Various models were built which included Random Forest, AdaBoost, Support Vector Machines, K-Nearest Neighbours, Naïve Bayes. In order to train the models, train-test split was done. Out of these models, the best performing model was Random Forest which achieved an AUC score of 73%.

2. Predicting and Evaluating the Popularity of Online News

   In the research, feature selection techniques such as mutual information, fisher criterion were used. As the dataset has a large number of dimensions, dimensionality reduction techniques (PCA) were applied but didn't provide any significant improvement in the models. Finally, top 20 features were selected and models such as Linear Regression, Logistic Regression, Support Vector Machines, Random Forest were implemented. The Random Forest was final model with 69% accuracy and 71% recall.

3. Online News Popularity Prediction

   The research used the top 20 features suggested in the previous paper. They also applied cfsSubsetEval evaluators to fetch appropriate features for model building. The models implemented included Random Forest, K-Nearest Neighbours, Logistic

Regression, Multilayer Perception (MLP). Among these models, Random Forest and MLP performed efficiently having f-score of 65%.

4. News Popularity Prediction with Ensemble Methods of Classification

   This paper followed before publication approach, where in the problem was converted to binary classification considering the threshold of 3395 with classes labelled as popular and unpopular. Recursive Feature Elimination technique was used to identify the relevant features that contribute towards popularity of articles. A total of 30 features were selected and models such as Naïve Bayes, Neural Network, Decision Tree, Random Forest, Gaussian and Support Vector Machine were implemented. The models Neural Network, Random Forest, Gaussian provided an accuracy of 79%.

5. Predicting the Popularity of Online News from Content Metadata

   A different way of predicting the popularity before publication was suggested in one of the articles, that is, their goal was to predict whether a news article may be shared or not by users as well as estimating the total count of shares. The models implemented were Gradient Boosting Machine (GBM) Regressor, Random Forest Regressor, GBM Classifier. A 5-fold cross-validation was used. The evaluation metric for regression was mean absolute percentage error (MAPE) and for classification AUC-ROC. The best AUC value obtained on Mashable news dataset was 74.5% and MAPE value of 69.42% using GBM.

6. Prediction & Evaluation of Online News Popularity using Machine Intelligence

   The research proposes that ensemble methods better predict the popularity of online articles. They determined the number of shares for each article. Here, in order to reduce the dimensions a dimensionality reduction method called as Linear Discriminant Analysis (LDA) was used. The data was split into varying ratios of train and test for model building. The models included LPBoost, AdaBoost, Random Forest and the comparison was based on these train-test split ratios. The AdaBoost model was the best model with accuracy 69% and F-score of 73%.

7. Predicting Popularity of Online Articles using Random Forest Regression

   This research involves predicting the number of shares before publication using various models such as Linear Regression, Random Forest Regressor, AdaBoost Regressor, Lasso and Ridge Regression. All these models were compared based on bias/variance/accuracy. The feature selection was based on the variance threshold method and random forest feature selection methods. The evaluation metric used was r2_score. Random Forest Regressor performed well by classifying 88.8% of the articles accurately as either popular or unpopular.

Dataset and Domain

3.1 Data Description:

- The dataset is Online News Popularity Prediction
- It consists of various attributes related to the articles published by Mashable over a period of 2 years from January 7, 2013 to January 7, 2015.
- Mashable is a well-known online news website founded in 2005. It covers all types of articles from travel to entertainment.
- A total of 39,644 articles were published in the span of 2 years. This data has been effectively scraped by researchers Fernandez, Vinagre and Cortez and donated to UCI Machine Learning Repository.
- There are total of 61 attributes. Out of which two are non - predictive (url, timedelta) and remaining 59 are numerical attributes.

3.2 Data Dictionary:

The attributes provided in the dataset can be grouped into different aspects as follows-

Aspects	Attributes
Words	Number of words of the title/content, Average word length, Rate of unique/non-stop words of contents
Links	Number of links, Number of links to other articles in Mashable
Digital Media	Number of images/videos
Publication Time	Day of the week/weekend
Keywords	Number of keywords, Worst/best/average keywords (shares)
Article category	Mashable data channels (bus, socmed, tech, world, lifestyle, entertainment)
NLP	Closeness to five LDA topics, Title/Text polarity/subjectivity, Rate and polarity of positive/negative words, Absolute subjectivity/polarity level
Target	Number of shares at Mashable

Table 1: Attributes of dataset

3.3 Pre-processing of Data:

- The dataset is almost clean, that is, it does not contain any null/missing values.
- The attribute names were corrected by removing the leading space, so that accessing the attributes would be easy.
- Also, some of the attribute names were not appropriate which were then renamed.

3.4 Project Justification:

- Project Statement

  To predict the popularity of online news articles before publication

- Complexity involved

  The original data is not provided, instead the statistics derived from the news articles on Mashable were used to create the dataset. The complex part was understanding of the given attributes and how they are calculated in order to relate them to the target variable.

- Project Outcome

  The outcome of the project is commercial as it would benefit the content writers of news publications or organizations to fine tune their article content so as to gain popularity.

Exploratory Data Analysis

Initially, the problem given was a regression problem to predict the 'number of shares' (target/independent variable) of the Mashable articles.

As most of the records had significant number of shares and just predicting the number of shares would not provide desired information or results regarding the popularity of the articles. So, in order to make the problem more precise it was converted from regression to classification problem.

The median of 'number of shares' was used to segregate the records into different classes which had a value of 1400. A new attribute named 'class' was created and based on a condition the labels were assigned to the records, this converted the problem into a classification problem.

Number of shares	Class
> 1400	popular (1)
< 1400	unpopular (0)

Table 2: Conditions for assigning labels

Considering target variable as both numerical (number of shares) and categorical (class), analysis of the data was done by studying the various aspects of the variables such as descriptive statistics, distribution of variables (Univariate and Bivariate Analysis). After analysis, some observations and interesting insights were noted.

4.1 Observations:

- Most of the attributes are of float datatype except for url (object) and shares (integer).
- The attributes are not highly correlated with target.
- There is high degree of skewness (right skewed) for each of the independent variable.
- Any independent variable is not linearly related to the target variable. Also, some kind of non-linearity exists.
- According to the correlation plot, a few independent variables are highly correlated with each other which indicates that multi-collinearity exists.
- Variables that show significant linear correlation seem to be heteroscedastic.
- The number of images in articles is more than the videos.
- Most of the articles are published on weekdays such as Tuesday, Wednesday and Thursday.

4.2 Insights:

1. Words/ Digital Media:
    a. The shares are high for the articles having a moderate number of words ranging from 7 to 15
    b. The articles which are not lengthy (382-2591 words) seem to gain maximum shares

c. Articles with number of images and videos ranging from 0 to 2 are likely to be shared.
d. Most (63%) news articles do not have videos, 24% articles have only 1 video, while the rest (13%) have more than 2 videos.
e. Most (46%) news articles have just 1 image each, 18% articles have no images, while the rest (36%) have more than 1 image.
f. There is no relationship between the number of images in an article and the number of times the article has been shared.
g. Maximum shared articles are the ones which have 19 words in their title
h. Articles which do not have any textual content seem to have images or videos.

Number of Articles	Title	Textual Content	Images	Videos	Shares
1181	yes	-	-	yes	yes
264	yes	-	yes	may/ may not	yes
716	yes	-	may/ may not	yes	yes
100	yes	-	yes	yes	yes
101	yes	-	-	-	yes
0	-	-	-	-	-

Table 3: Articles with no Textual Content

2. Links:
    a. A few links to other Mashable articles and other articles contributes towards good amount of shares
    b. Also, a significant drop in shares is observed if there are no links to other Mashable articles or other articles
    c. The articles having large number of links get very few shares
    d. It is observed that articles with no links to other articles including Mashable, still have shares.
    e. Articles with total of 4, 5 or 3 links constitutes about 8% of all the articles and have maximum shares
    f. There are articles with no links which have enough number of shares than the articles with exactly one link.

3. Keywords:
    a. Articles having number of keywords 3 or more are mostly shared
    b. Worst keywords (keywords with minimum shares): 58% of the news articles have (worst) keywords with -1 shares, 30% of the news articles have (worst) keywords with 4 shares, 11.7% of the news articles have (worst) keywords with 217 shares.
    c. There are 114 news articles which have just 1 worst keyword.
    d. There are 165 news articles having 1 best keyword.
    e. There are 130 news articles having 1 average keyword.

4. Article Category:
    a. The world category has highest number of articles but contributes only 19% of the shares and gained popularity of 34% within the category.
    b. Lifestyle and social media categories have least number of articles.
    c. Even though social media category has less articles, it has managed to get high popularity of 71% within the category.
    d. About 6134 articles might be of some other category which is not mentioned in the dataset. Since the data channel has been one-hot encoded, it seems that a column (unknown data channel) has been dropped. After research, we understand that the articles that have the most shares belong to the data channel (viral) that is dropped by one-hot encoding.
    e. From the given data channels, majority of the articles that have maximum shares are the ones that have positive sentiments and from an unknown data channel (viral data channel).
    f. Overall popularity of technology related articles is high about 27% as compared to popular articles of other categories

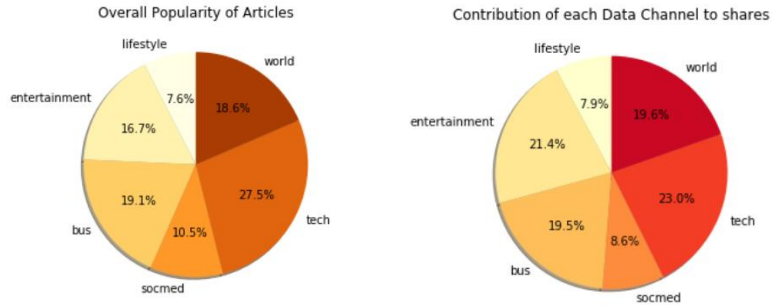

Figure 1: Article Category wise popularity and distribution of shares

Articles	Factual/ opinion (title and content)	Weekdays	Weekends	Primary Data Channel
18972	Opinion	yes	-	World
2582	Opinion	-	yes	World
2989	Factual	yes	-	Entertainment
561	Factual	-	yes	Entertainment
5	Opinion/ Factual	yes	-	-
0	Opinion/ Factual	-	yes	-

Table 4: Title and Content of articles with similar Subjectivity

Articles	Factual/opinion (title)	Factual/opinion (content)	Weekdays	Weekends	Primary Data Channel
5980	Factual	Opinion	yes	-	Entertainment, Tech
930	Factual	Opinion	-	yes	Tech
4337	Opinion	Factual	yes	-	World, Entertainment
622	Opinion	Factual	-	yes	Entertainment, World

Table 5: Title and Content of articles with dissimilar Subjectivity

5. Publication Time:
    a. The articles published on weekdays (Monday, Tuesday and Wednesday) contribute an equal number of shares of 15%.
    b. About 67% articles are popular among the articles that are published on weekends
    c. Articles of each data channel observed the least number of releases on weekends except Lifestyle
    d. The overall popularity of weekend articles is 15% among all the popular articles published on weekdays

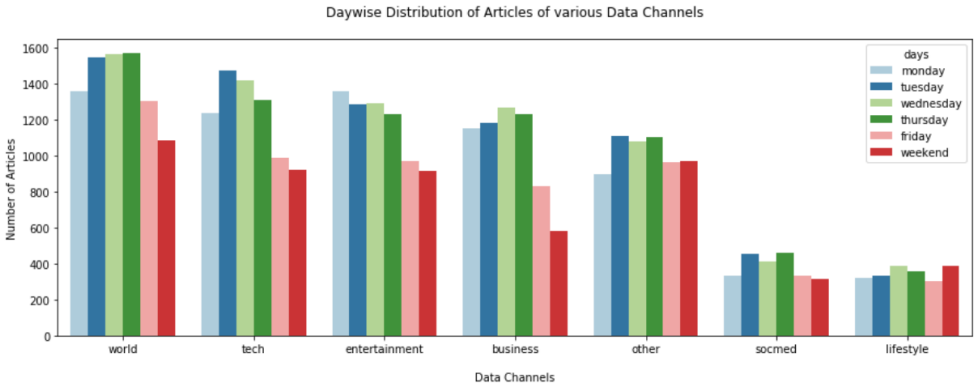

Figure 2: Distribution of articles day-wise

    e. The data channels world and entertainment observed a drastic increase in the number of articles published in the year 2014 as compared to 2013.
    f. Lifestyle and social media data channels have the highest average shares for both the years even though there is significant increase and decrease in average shares for the year 2014 respectively.
    g. The average shares for world data channel dropped in the year 2014 in spite of increase in the number of articles.
    h. We can say that, increasing the number of articles for a particular data channel does not contribute towards increase in number of shares or popularity.

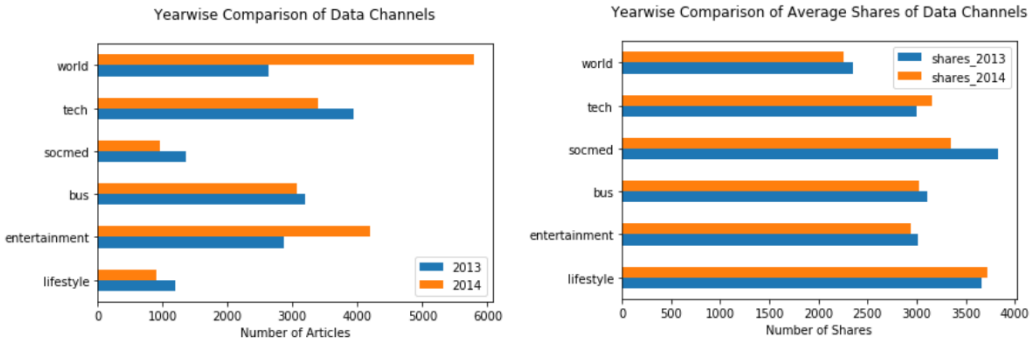

Figure 3: Yearly Comparison of Articles based on number of articles and shares

6. NLP:
    a. LDA Topics:
        i. Higher number of articles are related to the topic 4 whereas less articles are related to topic 1
        ii. The articles with article categories as business, technology, world, entertainment and other are closely related to topic 0, topic 4, topic 2, topic 1 and topic 3 respectively.
        iii. More than 10% of the articles having article categories as lifestyle, entertainment, social media and others are slightly related to topic 4, topic 3, topic 0 and topic 1 respectively.

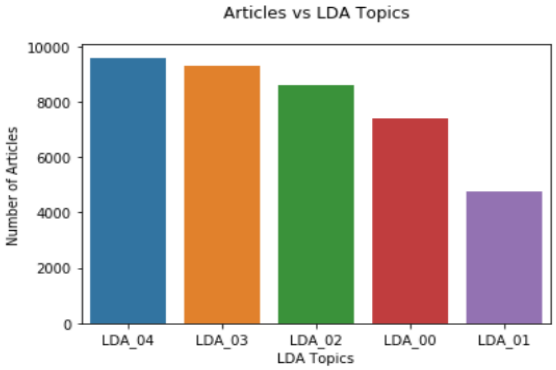

Figure 4: Number of Articles in each LDA topic

    b. Text/Title Subjectivity and Polarity:
        i. Most (71%) news articles are objective that is, factual rather than opinion based.
        ii. Most (89%) news articles have positive text sentiments, 8% of news articles have negative sentiments, 3% of news articles have neutral sentiments.
        iii. Many (45%) articles have factual titles irrespective of their content.
        iv. abs_title_subjectivity and abs_title_sentiment_polarity are scaled version of the original title_subjectivity and title_sentiment_polarity.
        v. Majority (50%) news articles have a neutral title, while 35% news articles have positive title polarity as compared to 89% of news articles having positive content polarity. This implies that most articles that have positive content may not have positive titles.
        vi. For few (10) articles that have a title and no content (textual, images, videos), links to other articles (Mashable and others), global subjectivity, sentiment polarity, positive/ negative polarity and keywords.
            - However, the title_subjectivity is between 0 to 0.5, which implies that these news articles are more of opinions than facts.
            - They also have title_sentiment_polarity between -0.02 to 0.3, so as to say that these news articles have neutral sentiments.
            - These articles have keywords (probably from the title and metadata)
            - These news articles have significant shares (4 to 5000).

vii. For the news articles which are highest shared (>80000)

- 49 of the news articles are objective (global_subjectivity<0.5)
- 39 of the news articles are subjective (global_subjectivity >=0.5)
- Articles that have maximum shares have positive sentiments/ positive polarity
- We cannot say that global subjectivity can be used to predict shares of news articles.

Articles	Sentiment polarity (title and content)	Week days	Weekends	Primary Data Channel
932	Negative	yes	-	World, Entertainment
166	Negative	-	yes	World, Entertainment
453	Neutral	yes	-	-
53	Neutral	-	yes	-
10942	Positive	yes	-	Tech, entertainment, business, world
1872	Positive	-	yes	Tech, entertainment, world

Table 5: Title and Content with similar Polarity

Articles	Sentiment polarity (title)	Sentiment polarity (content)	Weekdays	Weekends	Primary Data Channel
3983	Negative	Positive/ Neutral	yes	-	World, Entertainment
580	Negative	Positive/ Neutral	-	yes	World, Entertainment
17071	Neutral	Negative/ Positive	yes	-	World, Tech
2337	Neutral	Negative/ Positive	-	yes	World, Tech
959	Positive	Negative/ Neutral	yes	-	World, entertainment
171	Positive	Negative/ Neutral	-	yes	World, entertainment

Table 6: Title and Content with dissimilar Polarity

4.3 Statistical Significance of Variables:

The target variable is categorical, class (0 and 1)

Dependent Variable	Independent Variable	Statistical Test Applied
Categorical	Numerical	Mannwhitneyu test
Categorical	Categorical	Chi-square test

Table 7: Statistical Tests

4.3.1 Categorical vs Numerical:

- As most of the numerical independent variables didn't follow normal distribution, Mannwhitneyu test was performed.
- This test revealed that the independent variables 'avg_negative_polarity', 'min_negative_polarity', 'max_negative_polarity' and 'abs_title_subjectivity' were insignificant.

4.3.2 Categorical vs Categorical:

- The independent variables like data channels and weekdays were categorical in nature, Chi-square test was performed to check statistical significance.
- All the categories of data channels and weekdays turned out to be significant except 'weekday_is_friday' is insignificant.

4.4 Class Imbalance and its Treatment:

There are two classes, 1 (popular) and 0 (unpopular). Here, each class consists of almost equal number of records 50% and 49% respectively. Thus, this indicates that the dataset is balanced.

Hence, there is no problem of class imbalance and the techniques for handling any imbalanced dataset are not required.

Figure 5: Class Distribution

4.5 Scaling/Transformation:

The scaling of data is required to bring the data onto similar scale. Any transformations applied to the dataset resulted in null or inf values for some of the attributes due to which applying transformation was not so effective.

4.6 Feature Selection/Dimensionality Reduction:

From the above analysis, we can say that dimensionality reduction techniques like Principal Component Analysis (PCA) or feature selection techniques like Recursive Feature Elimination (RFE) could be applied in order to deal with multi-collinearity among the independent features and to reduce the number of dimensions from the dataset.