CAPSTONE PROJECT

FINAL 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 lifespan 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 an 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 recent years, there has been a lot of research done to predict the popularity of news articles published by Mashable. From the previous research 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 a 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 the 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 a 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 based on shares and LDA category (data source)

- 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
  - This will help news agencies understand which category (data source) the article with specific characteristics comes from and predict its popularity based on shares.

  - 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 numbers 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 variables 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 data type except for url (object) and shares (integer).
- The attributes are not highly correlated with the target.
- There is a high degree of skewness (right skewed) for each of the independent variables.
- Any independent variable is not linearly related to the target variable. Also, some kind of nonlinearity exists.
- According to the correlation plot, a few independent variables are highly correlated with each other which indicates that multicollinearity 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.

4.1.1 Check for Normality:

1. To check for Normality, Shapiro-Wilcox statistical test was adopted. It is a way of telling whether a random sample comes from a normal distribution.
2. Q-Q plot was used which compares two different distributions (Theoretical and sampled distribution). If the two sets of data came from the same distribution, the points will fall on a 45degree reference line.

3. The below is an example of the qq-plot to test for normality supplemented with Shapiro-Wilcox test.

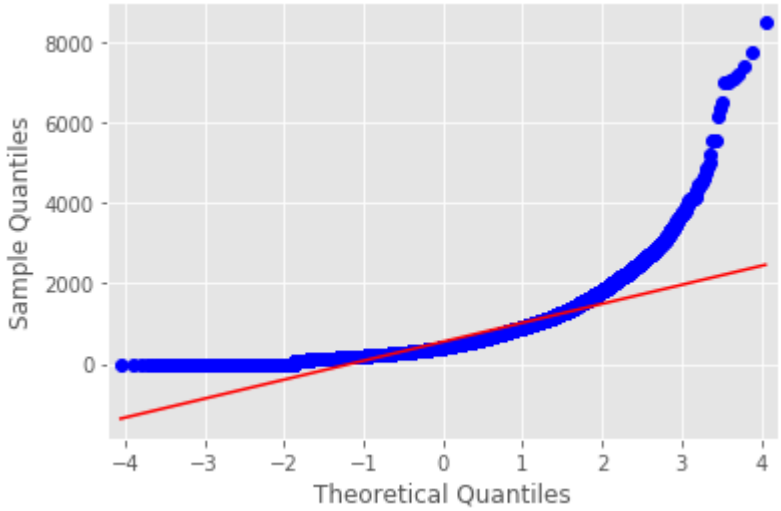

Figure 1: QQ plot for n_tokens_content

4. The Shapiro-Wilcox test returned a W statistic 0.7823435068130493 and a p-value of 0.0 which says that n_tokens_content does not come from a Normal distribution.

- After analysis of all the numeric attributes it was found that all the data showed non-normal distributions. Hence, non-parametric tests were the way to move forward.

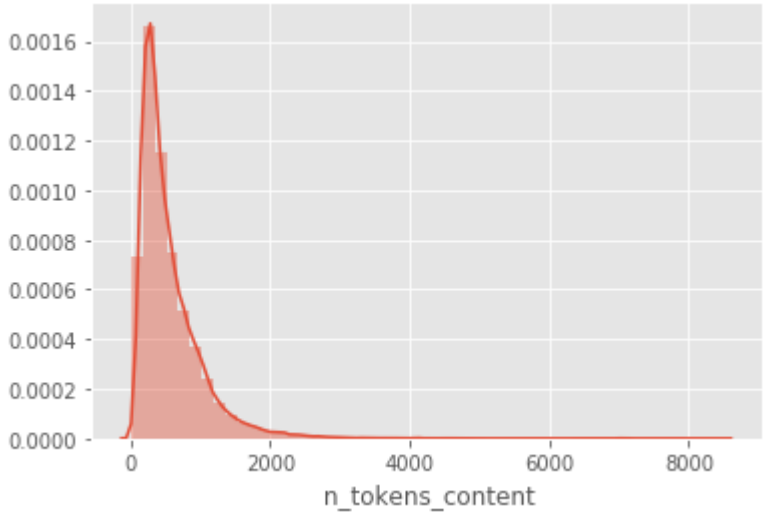

Figure 2: Distribution of n_tokens_content

- The shares variable has data distributed discreetly at the ends of the distribution.

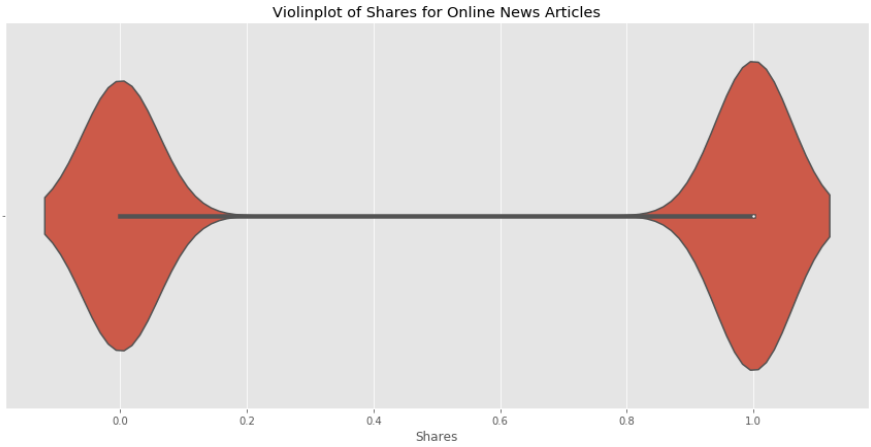

Figure 3: Distribution of shares variable

Multi Classification - Classification based on Unsupervised Approach

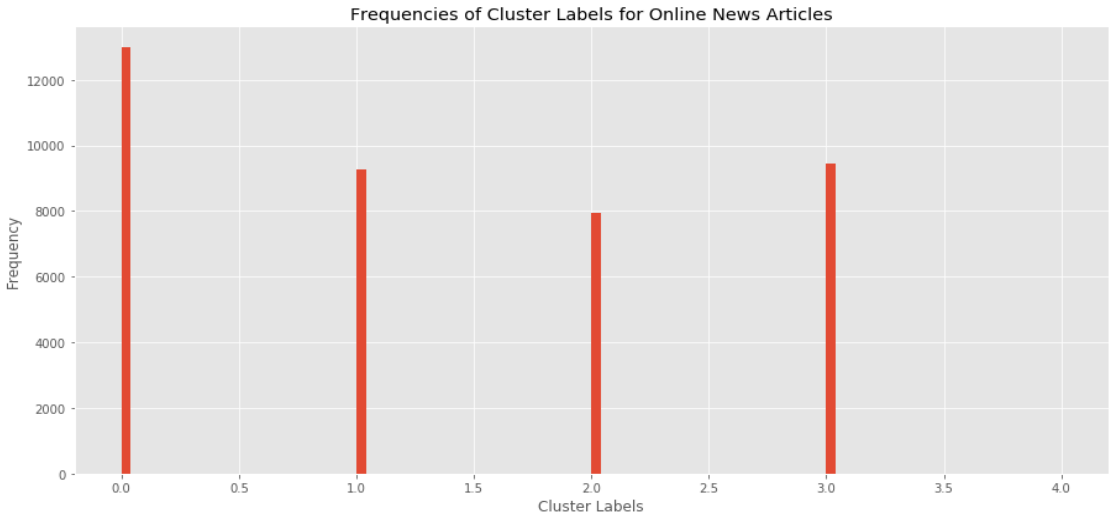

Figure 4: Frequency of cluster labels

- There are 5 clusters:
  - Cluster 0: News articles belonging to Entertainment data channel, LDA_03
  - Cluster 1: News articles belonging to World data channel, LDA_02
  - Cluster 2: News articles belonging to Business data channel, LDA_00
  - Cluster 3: News articles belonging to Tech data channel, LDA_02
  - Cluster 4: News articles belonging to No known LDA, data channel.

- As we can see that the frequency data from Cluster 0 is more than the rest of the clusters.
- This means that most of the data comes from LDA_03 which is related to the Entertainment data channel.
- The Cluster Labels variable has data distributed discretely throughout the distribution.

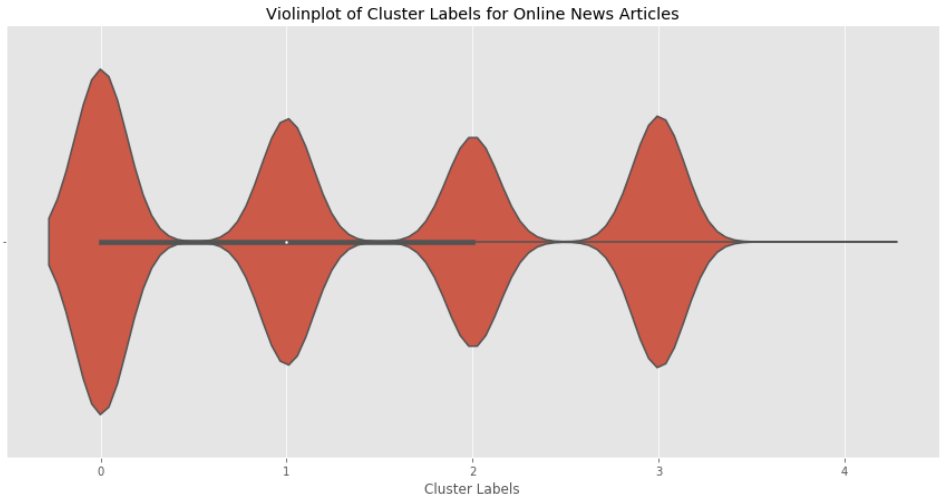

Figure 5: Distribution of records for each cluster

4.1.2 Distribution of target variable for Binary and Multi-Classification:

Binary Classification - Classification based on Shares:

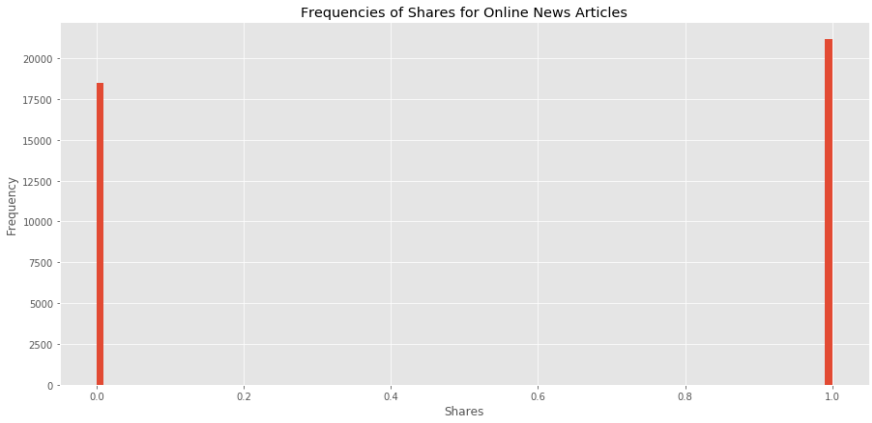

Figure 6: Distribution of shares

- There are two labels:
  - Label 0: Unpopular News Articles
  - Label 1: Popular News Articles
- As we can see that the frequency of popular articles (1) are more than that of unpopular articles (0).

4.1.3 Correlation with the target variable (Shares) for Binary Classification:

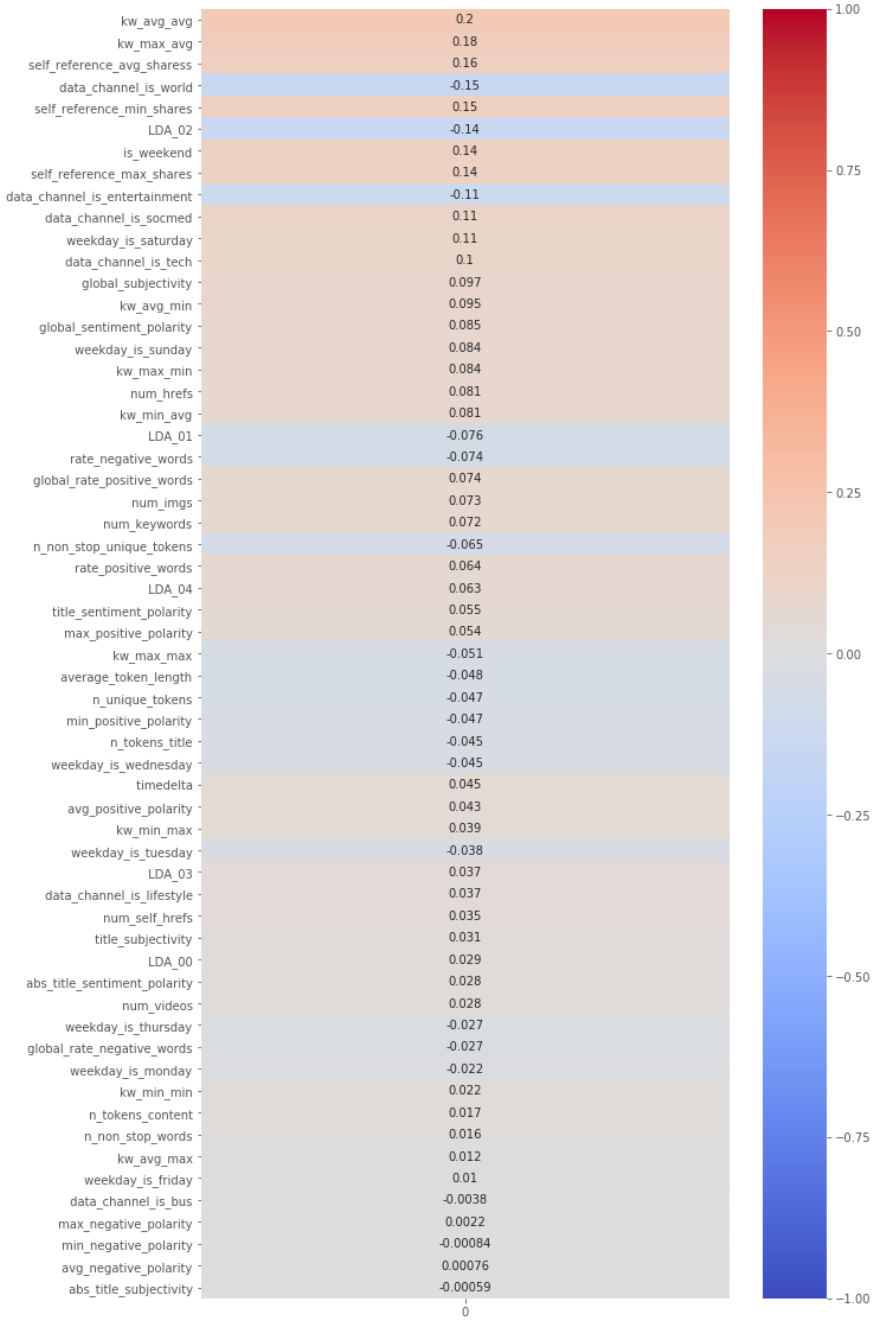

Figure 7: Correlation with shares

- The variables kw_max_avg, kw_avg_avg, self_reference_avg_shares, data_channel_is_world, self_reference_min_shares have the maximum correlation with the shares (target) variables as compared to the rest of the variables when measured using Spearman Correlation.
- However the strength of the correlation is not strong enough to prove that there is a significant relation with the shares (target) variable.

4.2 Insights:

1. Words/ Digital Media:
    a. The shares are high for the articles having a moderate number of words in the title ranging from 7 to 19
    b. The articles which are not lengthy (382-2591 words) seem to gain maximum shares
    c. Articles with a 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 titles.

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

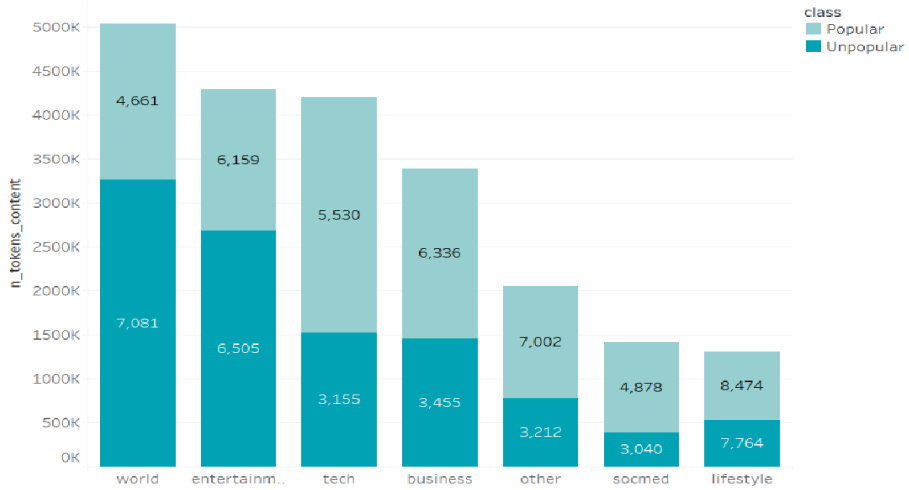

Figure 8: Popularity of articles based on their length

h. Lengthy news articles form the World data channel have a low popularity percentage (39.6%) and lifestyle articles are not that lengthy so are popular.
i. Social Media articles have highest popularity percent of 61%

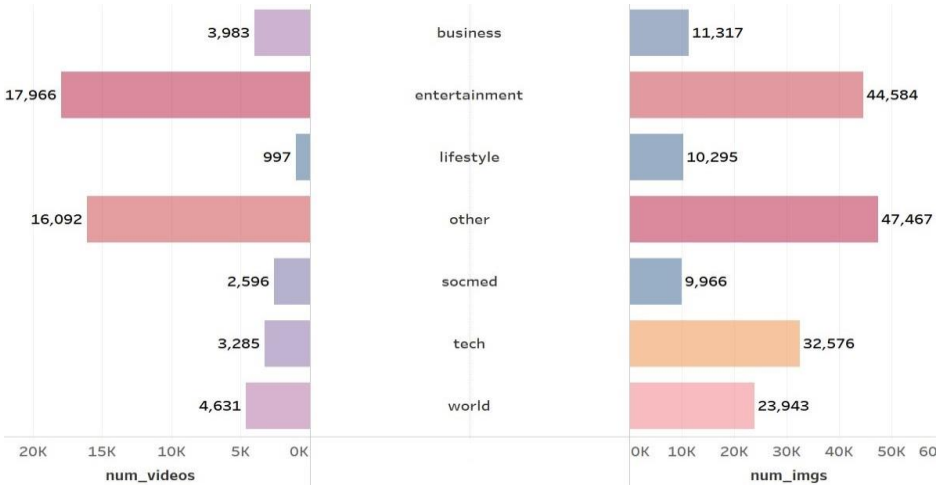

Figure 9: Number of Images and Videos

j. Most of the video articles are published on Tuesday and Wednesday as compared to weekends
k. The articles with images have approximately the same amount of releases on each day.
l. The articles that are published related to data channels others and entertainment include both videos and images.
m. We see that the articles having 8-13 words in the title and content between 382-2591 words have the maximum amount of shares but we don't have any statistical significance of the results.

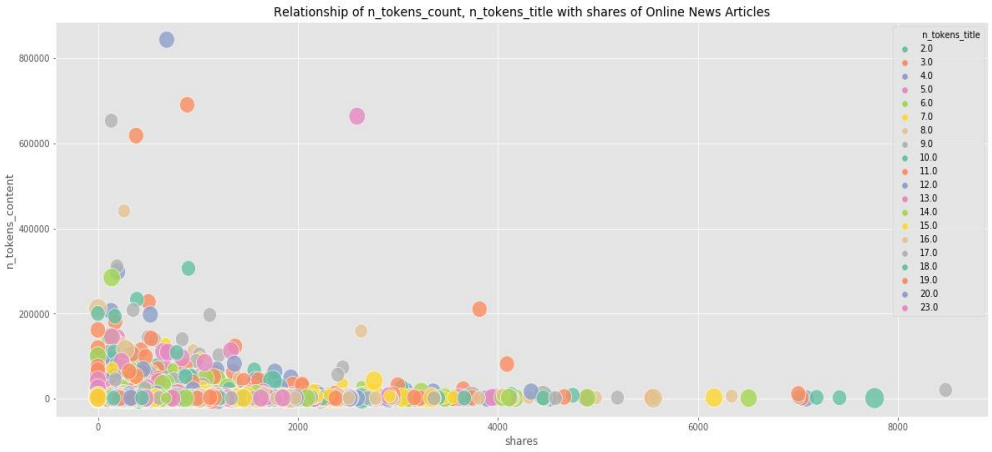

Figure 10: Relationship of number of tokens in content and title with shares

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 the 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 the least number of articles.
    c. Even though the social media category has fewer articles, it has managed to get a 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, the 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

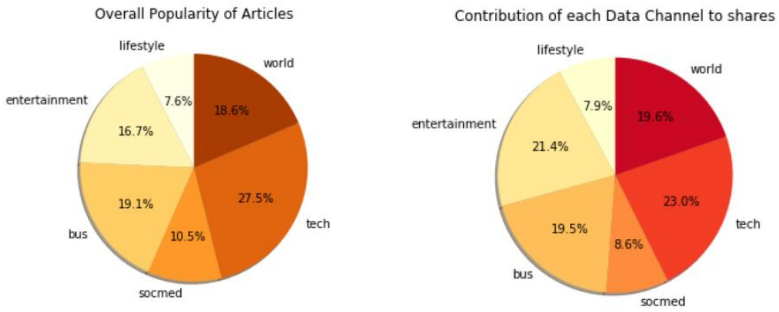

Figure 11: Article Category wise popularity and distribution of shares

g. Out of all data channels, the world channel has 21.26% shares, technology 18.53%, Entertainment 17.8%, business 15.79%, where social media 5.86 and lifestyle 5.29% have less shares.
    - Technology is most popular as compared to other channels 11% out of 49.35 after other, world and business 9.42% ,7.41 and 7.644 respectively lifestyle have popularity up to 3.03% which is very less.
    - World channels have the highest that is 13.85% unpopularity percentage and social media is 1.67% shares in unpopularity percentage.
    - Social media has the highest popularity ratio from their overall shares upto 71% (4.18/5.85)
    - World channels have the lowest popularity ratio from their shares 34.85%.

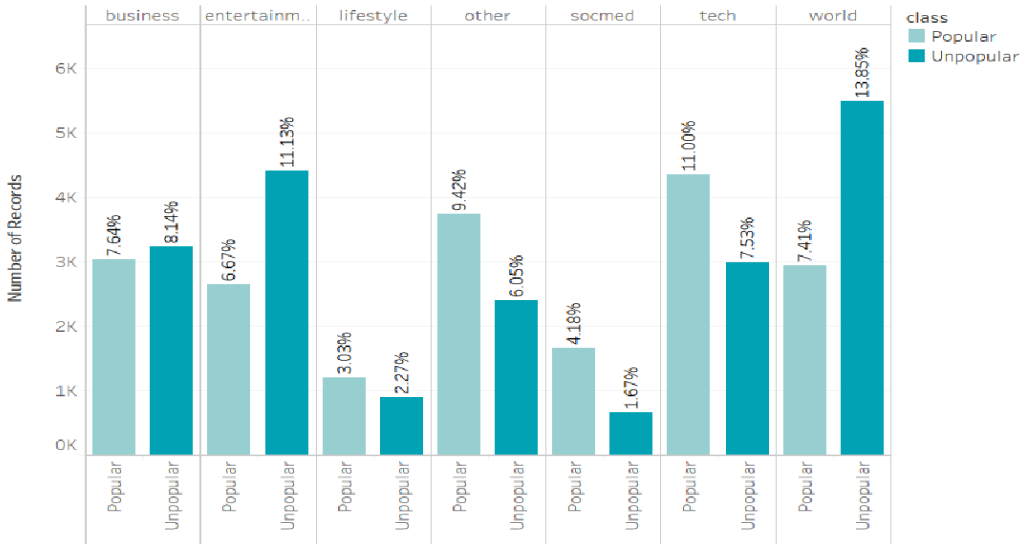

Figure 12: Popularity % of each data channel

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. Most of the articles published on Tuesday, Wednesday and Thursday are about 55.72% combined and the lowest article published on weekends is about 13.09%.
    d. Still weekends have the highest number of popularity percentage, about 67%.
    e. Wednesday have lowest number of popularity percent about 45.50%

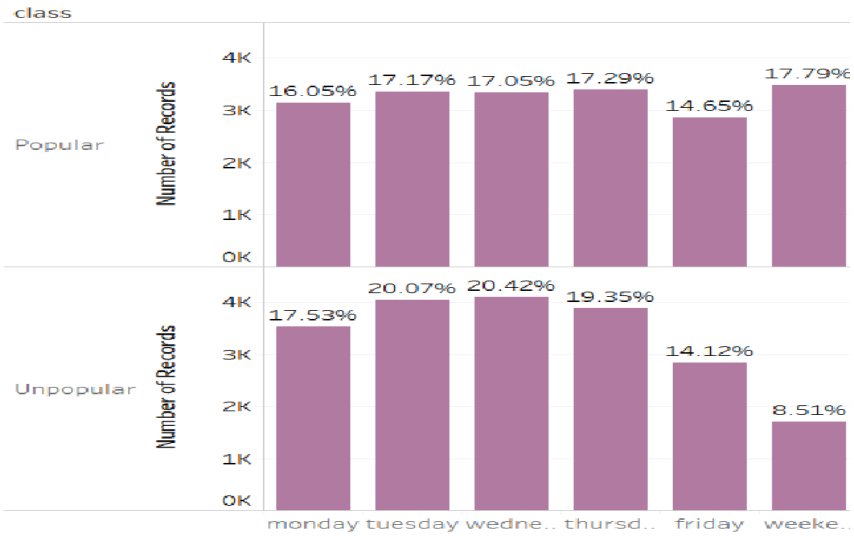

Figure 13: Day-wise Popularity of articles

    f. Articles of each data channel observed the least number of releases on weekends except Lifestyle
    g. The overall popularity of weekend articles is 15% among all the popular articles published on weekdays

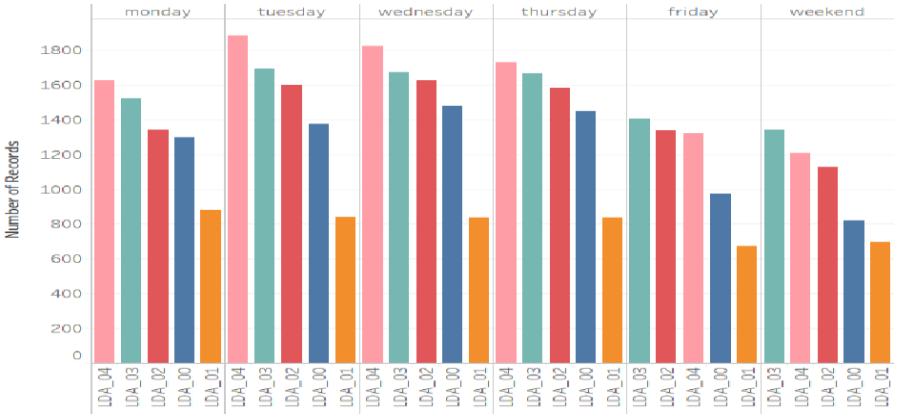

Figure 14: Distribution of articles day-wise

h. Distribution of articles based on LDA Topics

i. Monday: LDA_04 has 1623 articles, LDA_03 also has approximately the same number of shares (1518), LDA_01 had the lowest number of articles on Monday (879).

ii. Tuesday: LDA_04 has 1885 numbers of shares/articles, LDA_03 also has approximately the same number of shares (1690), LDA_01 has the lowest number of articles on Tuesday (840).

iii. Wednesday: LDA_04 has 1821 numbers of shares/articles, LDA_03 also has approximately the same number of shares (1672), LDA _01 had the lowest number of articles on Wednesday (837).

iv. Thursday: LDA_04 has 1730 numbers of shares/articles, LDA_03 also has approximately the same number of shares (1668), LDA_01 has the lowest number of articles on Thursday (835).

v. Friday: LDA_03 has 1409 numbers of shares/articles, LDA02 also has approximately the same number of shares (1334), LDA01 had the lowest number of articles on Friday (871).

vi. Weekends: LDA_03 has 1339 numbers of shares/articles, LDA_04 also has approximately the same number of shares (1207). LDA_01 had the lowest number of articles on weekends (697).

Figure 15: Year Wise Records

i. Out of 39644 shared news articles, 18199 articles (45.90%) have been published in 2013 and 21455 (54.10%) articles have been published in 2014.
j. From all the articles published, the percentage of popular articles is 46.61% in 2013 and 53.624% in 2014.

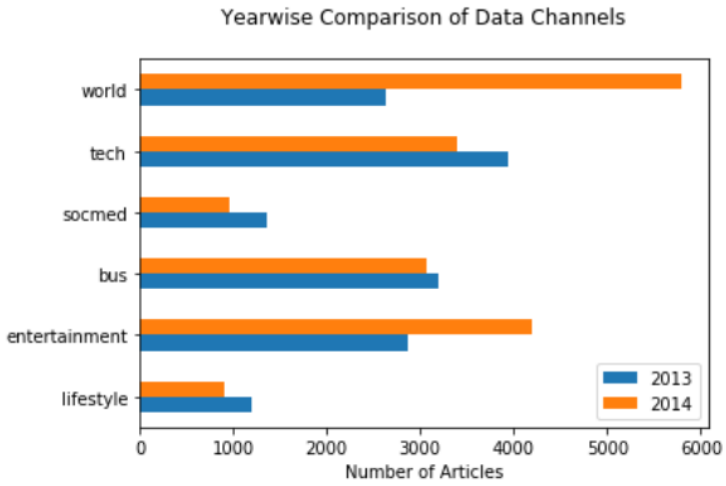

Figure 16: Yearly Comparison of Articles based on number of articles

k. The data channels world and entertainment observed a drastic increase in the number of articles published in the year 2014 as compared to 2013.
l. 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.

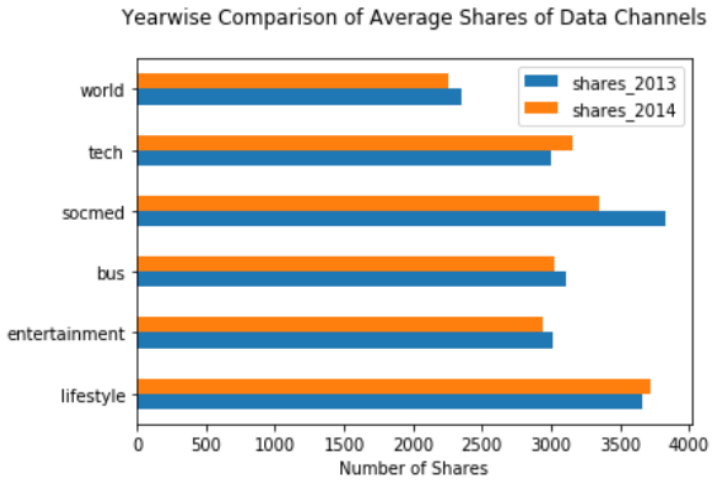

Figure 17: Yearly Comparison of Articles based on shares

m. The average shares for world data channels dropped in the year 2014 inspite of an increase in the number of articles.
n. We can say that increasing the number of articles for a particular data channel does not contribute towards an increase in number of shares or popularity.

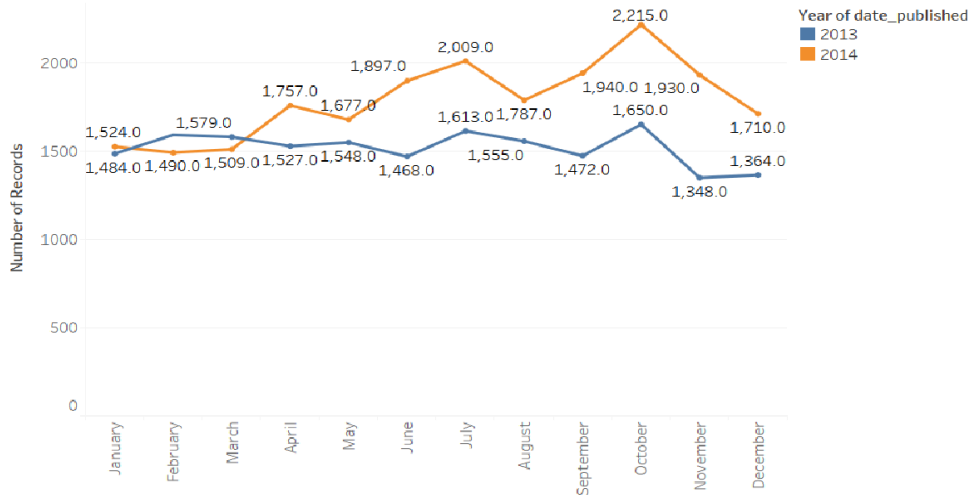

Figure 18: Month wise Records

o. October has the highest number of records in both the years, about 9.75% from the overall records and 9% in July.
p. January, February, December and March have approximately the same (7.75%) records, that is in these four months not many articles were published.
q. Still January has the highest popularity percent about 57.65%, March has 55% and the lowest popularity ratio is in October 44.5%.

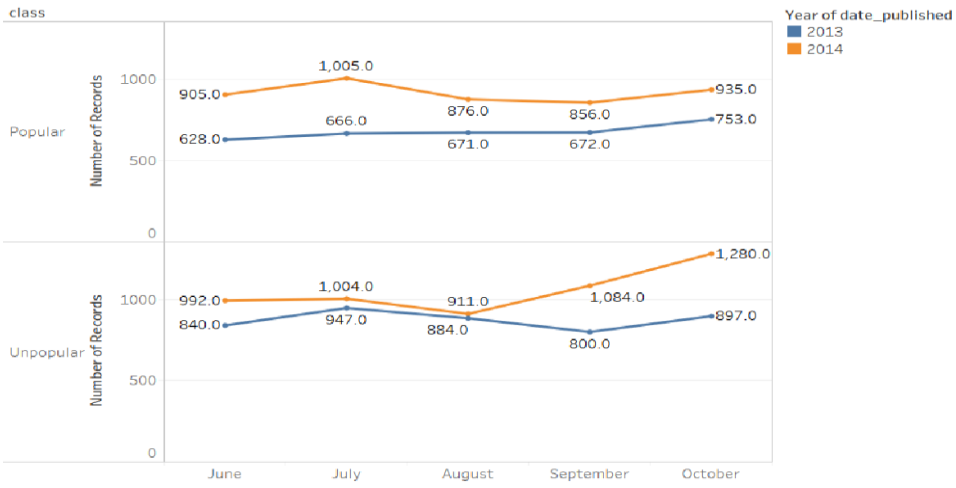

Figure 19: Top 5 months

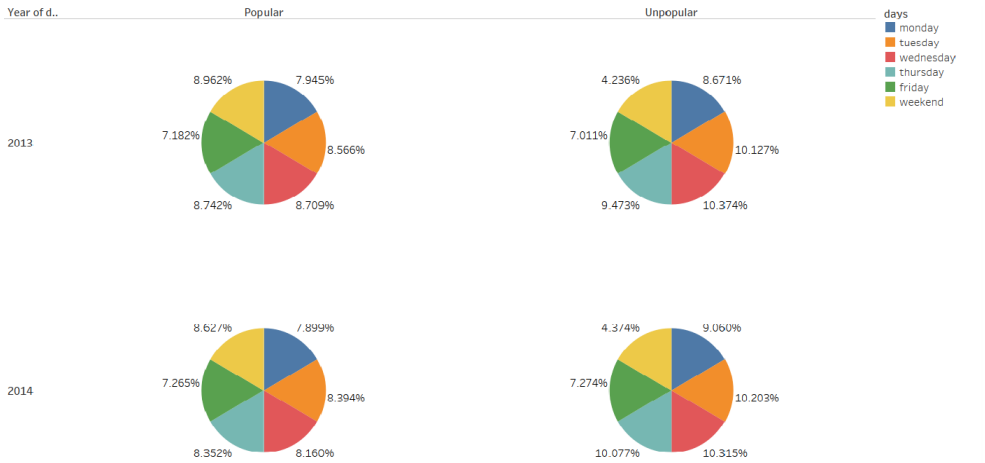

Figure 20: Weekday wise Popularity

r. For the year 2013,

Popular:

Out of 46.61% of popular articles, 8.96% articles are published on weekends.

All days have approximately the same percent of shares (7 to 8) in popularity.

Friday has the lowest popularity percent (7.182%).

Unpopular:

Out of 53.39% unpopular articles, 10.37% are published on Wednesday.

Tuesday and Thursday have 10.127% and 9.473% unpopularity percentages respectively.

Weekends have the lowest (4.12 %) unpopularity percentage.

s. For the year 2014,

Popular:

Out of 53.62% popular articles, 8.627% articles are published on weekends

All days have approximately the same percent of shares (7 to 8) in popularity.

Friday has the lowest popularity percent (7.265%)

Unpopular:

Out of 46.38% unpopular articles, 10.31% articles are published on Wednesday

Tuesday and Thursday have 10.20% and 10.07% unpopularity percentages respectively.

Weekends have lowest unpopularity percentage (4.374 %)

6. NLP:
    a. LDA Topics:
        i. Higher number of articles are related to the LDA_04 whereas less articles are related to LDA_01.
        ii. The articles with article categories as business, technology, world, entertainment and other are closely related to LDA_00, LDA_04, LDA_02, LDA_01 and LDA_03 respectively.
        iii. More than 10% of the articles having article categories as lifestyle, entertainment, social media and others are slightly related to LDA_04, LDA_03, LDA_00 and LDA_01 respectively.
        iv. Most of the topics are published in 2014.
        v. Most of the articles published have LDA_04 of about 24.18% (9585).

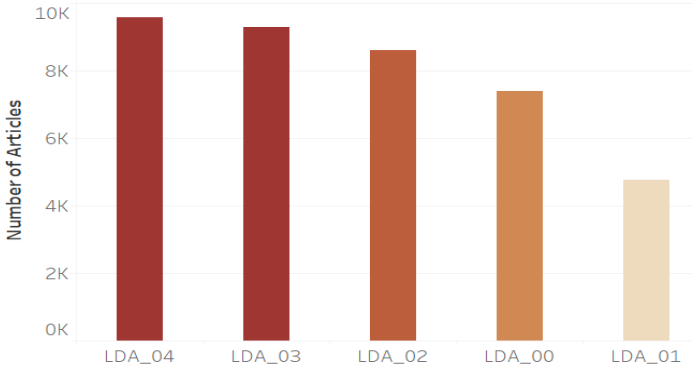

Figure 21: Number of Articles in each LDA topic

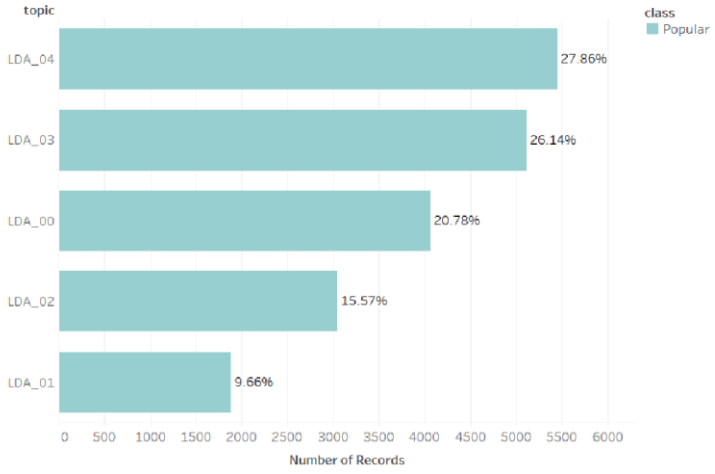

Figure 22: Data Channel vs LDA Topics

        vi. Few articles (4759) are related to LDA_01, about 12%.
        vii. LDA_04 topic have the most number of popular articles (5450) about 27.86% as compared to LDA_01 which has 9.66% (1889 articles)
        viii. Articles related to LDA_02 topics are highly unpopular where as LDA_04 topic have most high number of popular articles.

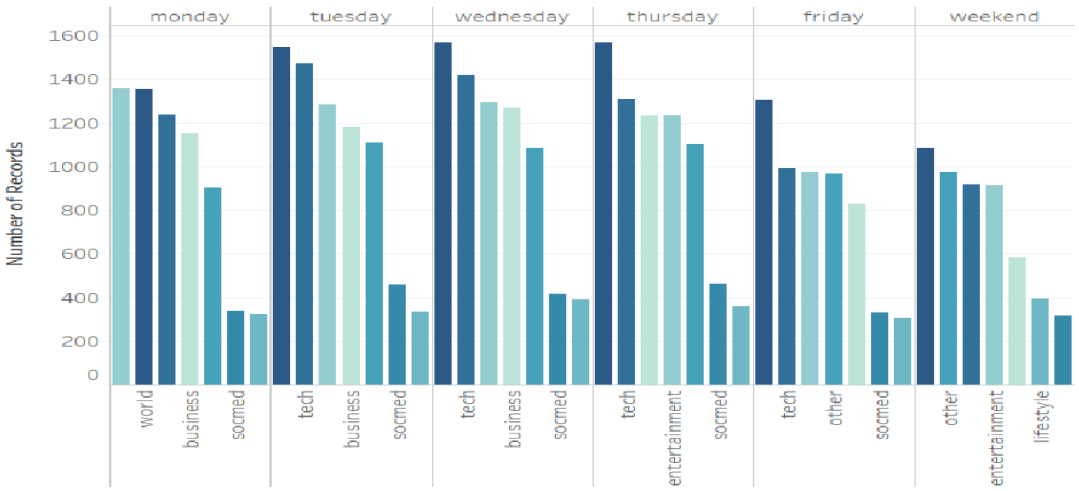

Figure 23: Days vs Data Channel

ix. For Monday, Entertainment has 1358 articles. World data channel also has approximately the same number of articles (1356). Lifestyle and social media data channels have the lowest number of articles on Monday, 322 and 337 respectively.

x. For Tuesday, Entertainment has 1358 articles. World data channel also has approximately the same number of articles (1356). Lifestyle and social media data channels have the lowest number of articles on Tuesday 334 and 458 respectively.

xi. For Wednesday, World data channel has 1565 articles and technology also has approximately the same number of articles (1417). Lifestyle and social media data channels have the lowest number of articles on Wednesday, 388 and 416 respectively.

xii. For Thursday, World data channel has 1569 articles. Lifestyle and social media data channels have the lowest number of articles on Thursday, 358 and 464 respectively.

xiii. For Friday, World data channel has 1305 articles. Lifestyle and social media data channels have the lowest number of articles on Friday, 305 and 322 respectively.

xiv. For Weekends, World data channel has 1096 articles. Lifestyle and social media data channels have the lowest number of articles on Tuesday, 392 and 317 respectively.

topic	data_ch	monday	tuesday	wednesday	thursday	friday	weekend
LDA_00	business	975.0	1,005.0	1,075.0	1,038.0	669.0	536.0
LDA_01	entertainment	730.0	661.0	670.0	655.0	517.0	399.0
LDA_02	world	1,120.0	1,306.0	1,342.0	1,313.0	1,122.0	927.0
LDA_03	other	805.0	977.0	966.0	984.0	865.0	700.0
LDA_04	tech	1,087.0	1,311.0	1,260.0	1,168.0	878.0	806.0

Figure 24: Data channels related to LDA

xv. For LDA_00, highest number of articles for business and mostly published on Wednesday and Thursday (1075 and 1038). Minimum number of articles from the world data channel and least published on weekends (9 articles only).

xvi. For LDA_01, highest number of articles of entertainment and mostly published on Monday (730). They have a minimum number of articles from lifestyle data channel and the least published on Monday (4 articles).

xvii. For LDA_02, the highest number of articles from the world data channel and mostly published on Wednesday and Thursday (1342 and 1314) respectively. They have a minimum number of articles from lifestyle data channel and least published on Monday (3 articles).

xviii. For LDA_03, the highest number of articles for others and mostly published on Tuesday, Wednesday and Thursday (977,962 and 984) respectively. They have a minimum number of articles from technology data channel and least published on weekends (5 articles only.

xix. For LDA_04, the highest number of articles from technology and mostly published on Tuesday (1311). They have a minimum number of articles from other channels and are least published on Thursday (8 articles).

b. Text/Title Subjectivity and Polarity:
   i. Most (71%) news articles are objective, 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. We see that (title_subjectivity and abs_title_subjectivity) and (title_sentiment_polarity and abs_title_sentiment_polarity) show non-monotonic relation.

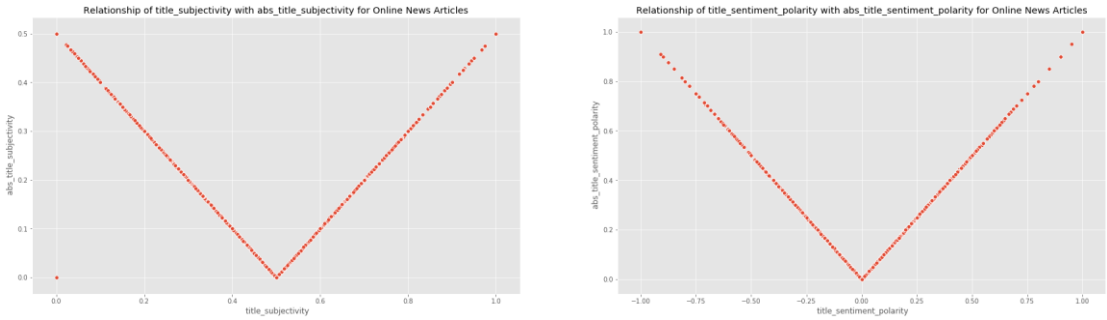

Figure 25: Non-Monotonic relationship between variables

iv. Many (45%) articles have factual titles irrespective of their content.

v. abs_title_subjectivity and abs_title_sentiment_polarity are scaled versions of the original title_subjectivity and title_sentiment_polarity.

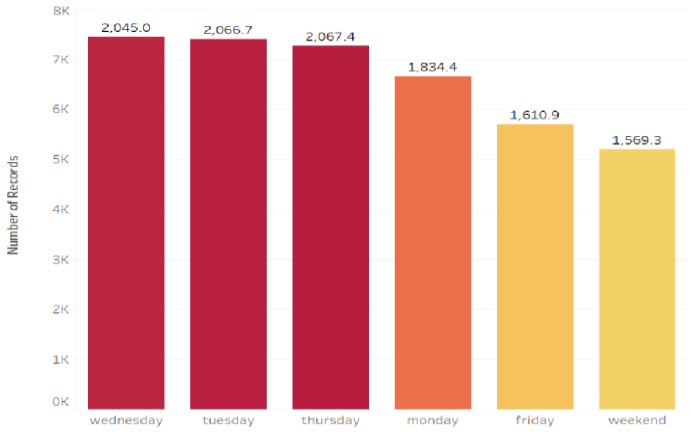

Figure 26: Day-wise subjectivity of articles

vi. Wednesday has a 2045 number of title subjectivity and weekends have a lowest 1539.3 number of title subjectivity.

vii. 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.

viii. For a 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).

ix. 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 6: 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 7: Title and Content with dissimilar Polarity

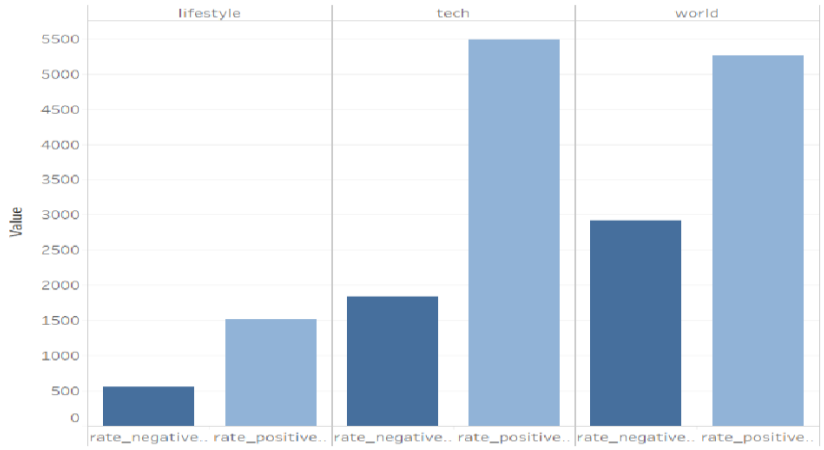

Figure 27: Data Channel wise rate of positive and negative words

x. World has the highest number of both positive and negative word rates. (Positive word 5253 and negative word 2915)

xi. Technology has the highest number of positive words. (Positive words is 5484 and negative words is 1841)

xii. Lifestyle has the lowest number of both positive and negative word rates. (Positive words 1517 and negative words 560)

Figure 28: LDA topic wise rate of positive and negative words

xiii. LDA_01 has less number of both positive and negative words

xiv. LDA_04 has the most number of positive words

xv. LDA_02 has the most number of negative words

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 8: Statistical Tests

4.3.1 Categorical vs Numerical:

- As most of the numerical independent variables didn't follow normal distribution, Mann Whitney U 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:

Binary Classification - Classification based on Shares:

There are two classes, 1 (popular) and 0 (unpopular). From the chart ,we can see that 50.65% online articles are not popular and 49.35% is popular 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.

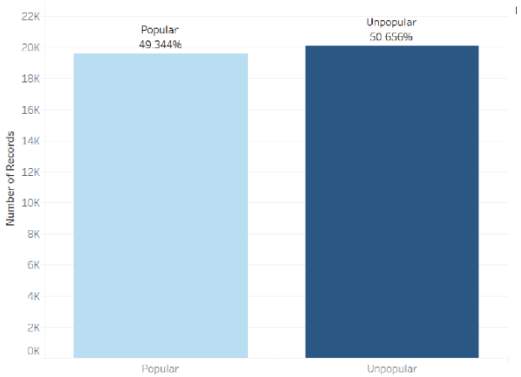

Figure 29: Class Distribution

Multi Classification - Classification based on Unsupervised Approach:

The target class labels (cluster labels) after KMeans Unsupervised approach were imbalanced. So, they were addressed using SMOTE analysis..

Figure 30: Class Imbalance

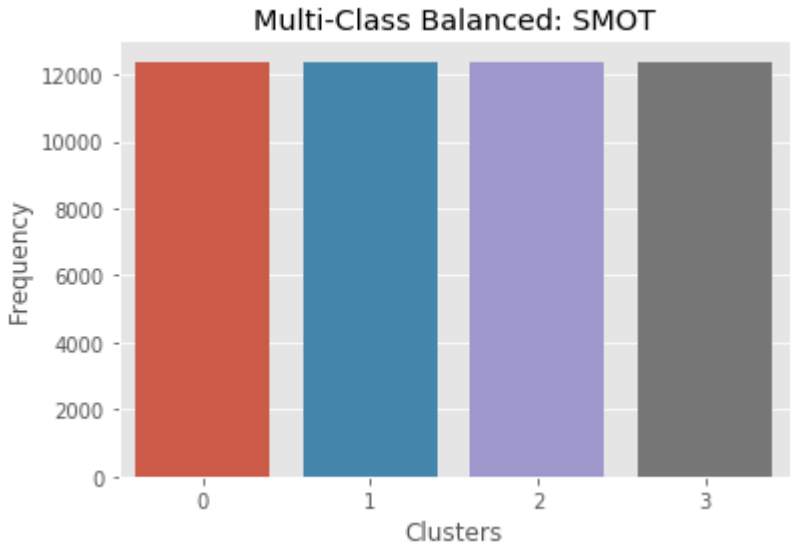

Figure 31: Balanced Class Labels using SMOTE

4.5 Scaling/Transformation:

- The scaling of data is required to bring the data onto a similar scale.
- As our data is mostly right skewed, to normalize it different transformations were applied on the given data.
- We tried applying log transformation but it gives infinity and null values so we decided to drop that idea. Further, square root and cube root transformations were

also applied and built a Logistic Regression model on top of that but it gave the same result as our base model.

- For this dataset applying transformation was not so effective so we decided to not go with any of the transformations for our final modelling part.

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 multicollinearity among the independent features and to reduce the number of dimensions from the dataset.

4.7 Outlier Detection and Treatment

The IQR method was used for detecting the outliers. It resulted in a large number of outliers that is about half of the data. So, removing those would have led to loss of some valuable information. Due to lack of domain knowledge, we decided not to remove or treat these outliers.

4.7.1 Outlier Detection:

Use of Boxplots with Strip plots to detect outliers and its concentration

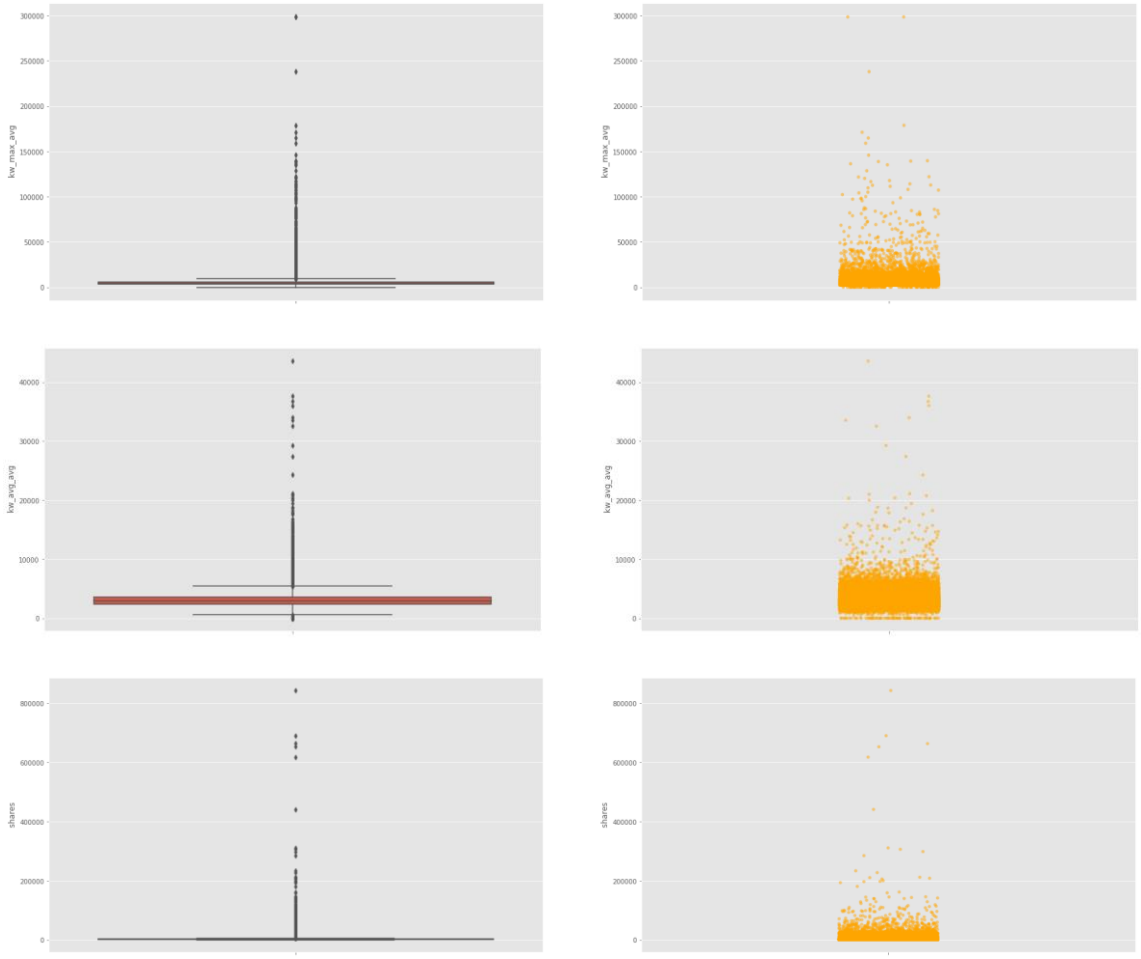

Figure 32: Outlier Detection

- The above figures show a combination of boxplots and strip plots for variables kw_max_avg, kw_avg_avg, shares to detect outliers as well as to measure the concentration of data.
- The variables chosen have a Spearman correlation greater than 0.2 and less than -0.2.

4.7.2 Outlier Treatment:

Z-Score: Z-Score was used to remove outliers which reduced the records from 39644 to 21009, which results in 18,635 records of usable data. However, this approach is not suitable as we lose a lot of relevant data.

Interquartile Range (IQR): Outlier treatment using IQR method however seems unviable as all the records were removed.

Transformation: The above outlier treatment approaches failed to provide a valuable solution to outliers so transformation was trained out.

The total outliers after Original transformation are: 191234
The total outliers after Squareroot transformation are: 151923
The total outliers after Logarithm transformation are: 95147
The total outliers after Sin transformation are: 159486
The total outliers after Cosine transformation are: 174622

Out of the following approaches Logarithmic transformation seemed to be more suitable however it results in inverse and NAN values. Hence, couldn't be used.

DBSCAN: Unsupervised Learning using Density-based spatial clustering of applications with noise (DBSCAN) was used as it is an effective approach for outlier treatment. However, most of the data was labelled as noise. Hence, this is not a suitable approach to handle outliers.

Out of all the approaches adopted none of the approaches seemed suitable for the dataset and hence modelling was done with outliers as the extreme values (outliers) proved to be relevant after further analysis.

Feature Engineering Techniques

A way of identifying or extracting a subset of features from the given data, transforming them into a format which is suitable for machine learning algorithms.

In this project, the methods such as Backward Elimination, Recursive Feature Elimination and Feature Importance were applied in order to get the suitable features.

1. Backward Elimination

This technique iteratively removes the features one by one based on the p-value (alpha = 0.05) till the performance of the model improves. After applying this technique to the given dataset, 42 features were selected.

Binary Classification - Classification based on Shares:

```
['n_tokens_content', 'n_non_stop_unique_tokens', 'num_hrefs', 'num_self_hrefs', 'average_token_length', 'num_keywords', 'data_c
hannel_is_lifestyle', 'data_channel_is_entertainment', 'data_channel_is_bus', 'data_channel_is_socmed', 'data_channel_is_tech',
'kw_min_min', 'kw_max_min', 'kw_avg_min', 'kw_min_max', 'kw_max_max', 'kw_avg_max', 'kw_min_avg', 'kw_max_avg', 'kw_avg_avg',
'self_reference_avg_shares', 'weekday_is_monday', 'weekday_is_tuesday', 'weekday_is_wednesday', 'weekday_is_thursday', 'weekday
_is_friday', 'weekday_is_saturday', 'weekday_is_sunday', 'is_weekend', 'LDA_00', 'LDA_01', 'LDA_02', 'LDA_03', 'LDA_04', 'globa
l_subjectivity', 'global_rate_positive_words', 'global_rate_negative_words', 'rate_positive_words', 'min_positive_polarity', 't
itle_subjectivity', 'title_sentiment_polarity', 'abs_title_subjectivity']
```

Figure 33: 42 Optimal features

2. Recursive Feature Elimination

In order to find the optimal number of features to train the model, Recursive Feature Elimination (RFE) technique was applied. However, for more precise results, RFE with cross-validation (RFECV) was used for feature selection.

Binary Classification - Classification based on Shares:

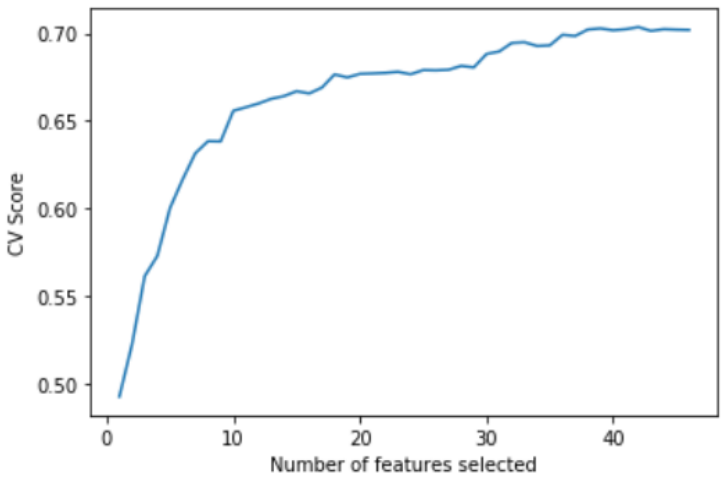

Figure 34: Optimal Number of Features

For RFECV, the parameters provided include random forest as an estimator with default values, cv as 5 and scoring as roc_auc. Then, a fit was applied on independent features excluding the insignificant ones based on statistical tests. This resulted in 42 best features for the classification model.

```
['n_tokens_title', 'n_tokens_content', 'n_unique_tokens', 'num_hrefs', 'num_self_hrefs', 'num_imgs', 'num_videos', 'average_tok
en_length', 'num_keywords', 'data_channel_is_entertainment', 'data_channel_is_socmed', 'data_channel_is_tech', 'data_channel_is
_world', 'kw_min_min', 'kw_max_min', 'kw_min_max', 'kw_avg_max', 'kw_min_avg', 'kw_max_avg', 'self_reference_min_shares', 'self
_reference_max_shares', 'weekday_is_tuesday', 'weekday_is_wednesday', 'weekday_is_thursday', 'weekday_is_saturday', 'is_weeken
d', 'LDA_00', 'LDA_01', 'LDA_03', 'LDA_04', 'global_subjectivity', 'global_sentiment_polarity', 'global_rate_positive_words',
'global_rate_negative_words', 'rate_positive_words', 'rate_negative_words', 'avg_positive_polarity', 'min_positive_polarity',
'max_positive_polarity', 'title_subjectivity', 'title_sentiment_polarity', 'abs_title_sentiment_polarity']
```

Figure 35: Features selected by RFECV

Multi Classification - Classification based on Unsupervised Approach:

To estimate the optimal attributes for Recursive Feature Selection (RFE) a random forest used was derived for a hyper-parameter tuned decision tree resulting with a test accuracy of 0.96 and 30 as the optimal number of attributes.

[' timedelta', ' n_tokens_content', ' n_unique_tokens', ' n_non_stop_unique_tokens', ' num_imgs', ' average_token_length', ' data_channel_is_lifestyle', ' data_channel_is_entertainment', ' data_channel_is_bus', ' data_channel_is_socmed', ' data_channel_is_tech', ' data_channel_is_world', ' kw_avg_max', ' kw_max_avg', ' kw_avg_avg', ' LDA_00', ' LDA_01', ' LDA_02',' LDA_03', ' LDA_04', ' global_subjectivity', ' global_sentiment_polarity', ' global_rate_positive_words', ' global_rate_negative_words', ' rate_positive_words', ' rate_negative_words', ' avg_positive_polarity', ' min_positive_polarity', ' max_positive_polarity', ' avg_negative_polarity']

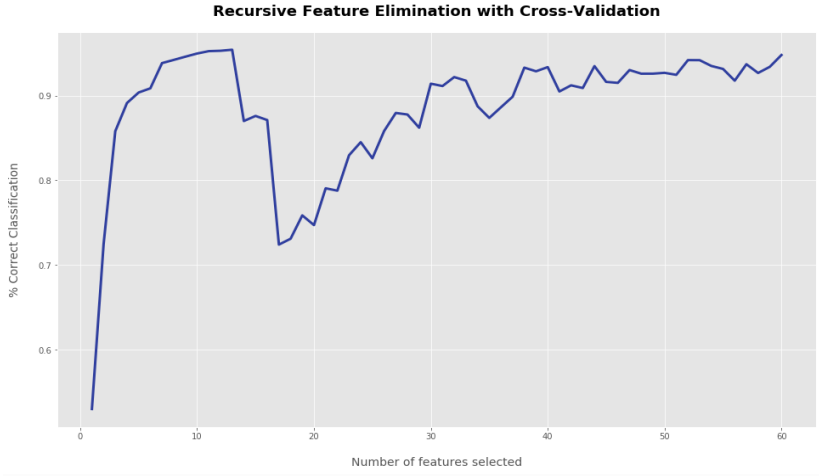

Figure 36: Optimal Number of Features by RFECV

For a more reliable and precise result, Recursive Feature Elimination with Cross Validation (RFECV) was adopted using Stratified Sampling on a random forest derived for a hyper-parameter tuned decision tree resulting in 13 as the optimal number of attributes and a test accuracy of 0.96.

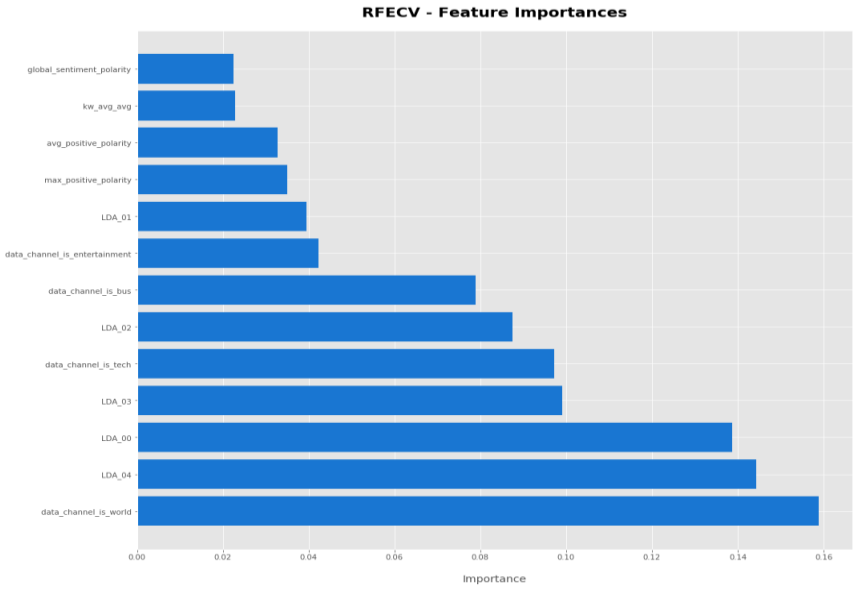

Figure 37: Ranking of Features based on Importance by RFECV

3. Feature Importance

A class of techniques for assigning scores to input features of a predictive model that indicates the relative importance of each feature when making a prediction.

In our project, this technique has been used for reducing the number of input features. For algorithms like decision tree or ensemble techniques, the property of providing importance scores is available after fitting the model which could be accessed by using the feature_importances_ attribute.

Binary Classification - Classification based on Shares:

Here, different subsets of features were selected (20, 25, 30, 40) from the tuned models after performing RFE and were evaluated. From these different models, the models with 20 features each were selected.

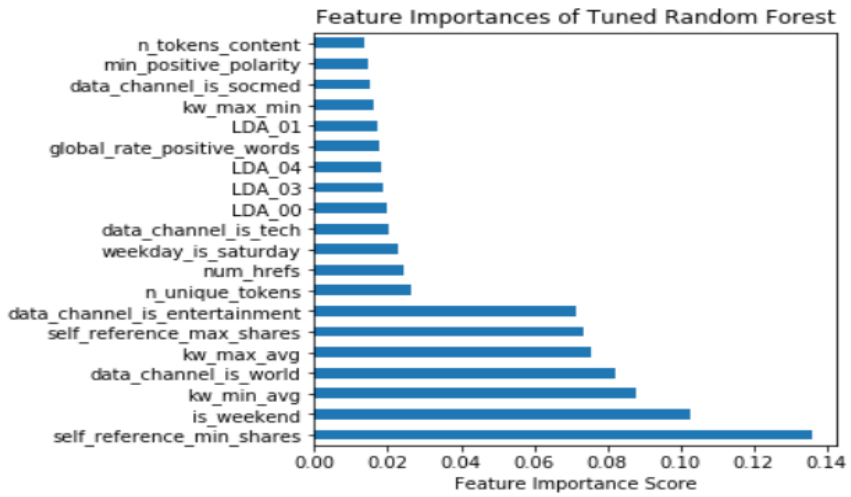

Figure 38: Top 20 features

Multi Classification - Classification based on Unsupervised Approach

To estimate the optimal attributes for Feature Importance a random forest used was derived for a hyper-parameter tuned decision tree resulting in 17 optimal attributes and a test accuracy of 0.96.

[' data_channel_is_world', ' data_channel_is_tech',' data_channel_is_bus', ' LDA_00', ' LDA_04', ' LDA_03', ' LDA_02', ' data_channel_is_entertainment', ' LDA_01', ' kw_avg_avg', ' kw_max_avg', ' data_channel_is_socmed', ' global_rate_positive_words',' kw_avg_max', ' data_channel_is_lifestyle', ' average_token_length',' avg_negative_polarity']

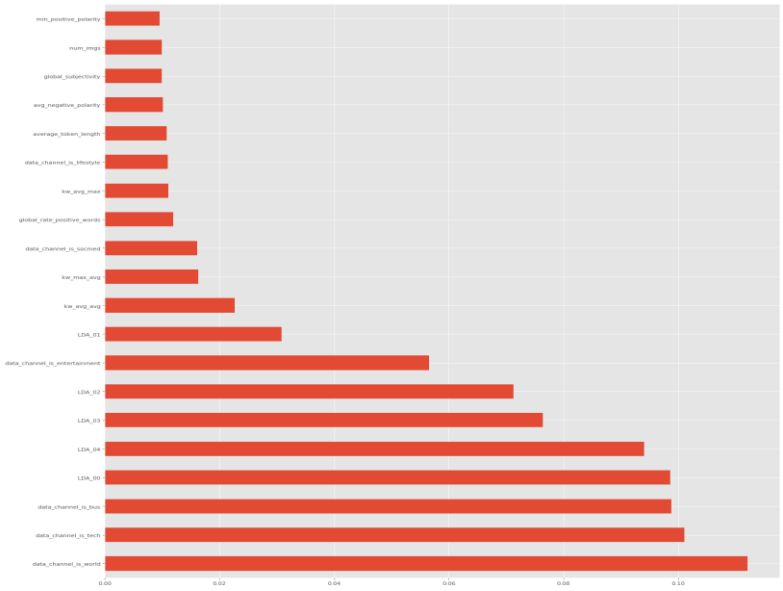

Figure 39: Ranking of top 20 features selected by Feature Importance

4. SelectKBest

Multi Classification - Classification based on Unsupervised Approach

SelectKBest follows a Univariate approach which uses statistical tests for selecting attributes to be used for modelling. We have used 'Chi-square' and 'mutual-information' as the statistical test for SelectKBest. The reason is that since the data is non-parametric, hence there is a need to use statistical tests which have non-parametricity as an assumption.

To estimate the optimal attributes for SelectKBest (chi-square) a random forest used was derived for a hyper-parameter tuned decision tree resulting in 17 optimal attributes and a test accuracy of 0.95

Features selected by SelectKBest (chi-square):

[' timedelta', ' n_unique_tokens', ' n_non_stop_words', ' n_non_stop_unique_tokens', ' num_videos', ' data_channel_is_lifestyle', ' data_channel_is_entertainment', ' data_channel_is_bus', ' data_channel_is_socmed', ' data_channel_is_tech', ' data_channel_is_world', ' LDA_00', ' LDA_01', ' LDA_02', ' LDA_03', ' LDA_04', ' rate_positive_words']

To estimate the optimal attributes for SelectKBest (mutual-information) a random forest used was derived for a hyper-parameter tuned decision tree resulting in 17 optimal attributes and a test accuracy of 0.95

Features selected by SelectKBest (mutual-information):
[' data_channel_is_entertainment', ' data_channel_is_bus', ' data_channel_is_tech', ' data_channel_is_world', ' kw_max_min', ' kw_avg_max', ' kw_max_avg', ' kw_avg_avg',' self_reference_max_shares', ' LDA_00', ' LDA_01', ' LDA_02', ' LDA_03', ' LDA_04', ' global_subjectivity', ' global_sentiment_polarity', ' rate_positive_words']

5. Correlation Matrix Analysis

Multi Classification - Classification based on Unsupervised Approach

Correlation Matrix Analysis uses “Spearman correlation coefficient” which is used to measure the correlation between attributes following non-normal distribution.

After using this approach we find that ' data_channel_is_entertainment' , ' LDA_03' , ' LDA_04' , ' data_channel_is_tech' (4 features) have correlation >0.5 with the target variable.

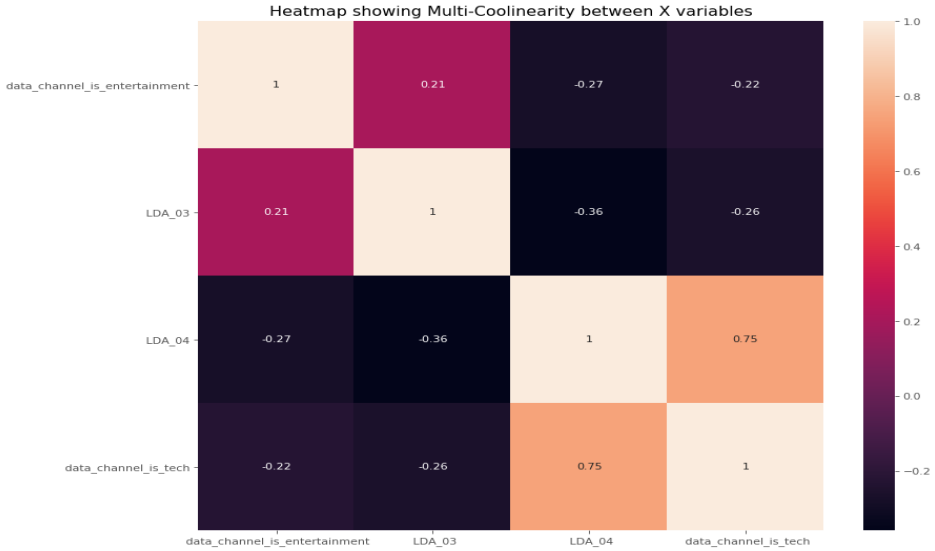

Figure 40: Correlation Matrix

However, multicollinearity seems to be weak among X features. There seems to be a relation between LDA_04 AND data_channel_is_tech, hence we would have to drop data_channel_is_tech.

Thus we get 3 features: ' data_channel_is_entertainment' , ' LDA_03' , ' LDA_04'

This is however insufficient to classify the target variable. Hence, it is not a viable approach.

Principal Component Analysis

A dataset with a large number of attributes may introduce multicollinearity and the model built with these attributes would be highly complex which may affect the model's performance. These issues of the dataset can be handled by dimensionality reduction techniques like Principal Component Analysis (PCA).

This Mashable dataset consists of a large number of attributes (61) and a slight multicollinearity also exists. In order to reduce the complexity and deal with multicollinearity, PCA was applied to reduce the number of attributes used for building the model. As this is a popularity prediction problem, for the results to be more accurate and to capture maximum variability within the data, the variance explained by the principal components was set to 95%.

Binary Classification - Classification based on Shares:

Scaled X data (excluding 'url' and target: 'shares') was used as input to PCA which resulted in 35 principal components.

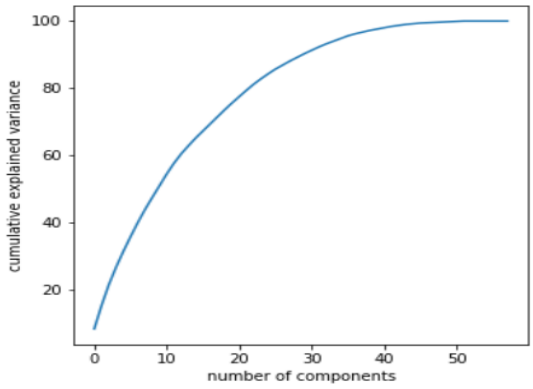

Figure 41: Optimal Number of PC's

Multi Classification - Classification based on Unsupervised Approach

Scaled data including target: 'shares' (excluding 'url') was used as input to PCA which resulted in 38 principal components. This would further be used for KMeans algorithm to cluster the data.

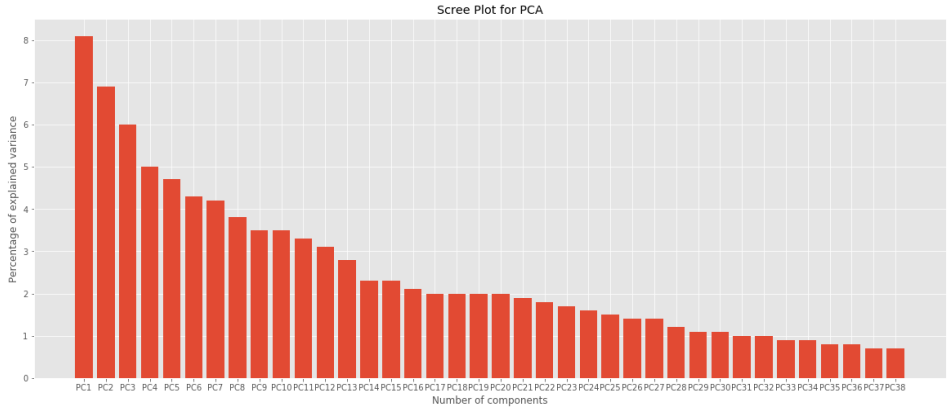

Figure 42: Number of PC's

Models and its Implementation

Based on the previous research on this dataset and from the results obtained from the exploratory data analysis, we can say that the non-linear models such as classification models can be effective for predicting the popularity of news articles.

7.1 Approaches for Model Building

Two approaches for model building can be taken into consideration which can be defined as follows:

- Using supervised learning approach, building binary classification models
- Using unsupervised learning approach, building multi-class classification models

7.1.1 Supervised Learning Approach:

In this approach, labels can be assigned to the records of the dataset as 0 and 1. This converts the problem into a binary classification problem. On this newly formed dataset, models can be built.

7.1.2 Unsupervised Learning Approach:

In this approach, it was assumed that the target variable is not available. Further, using clustering algorithms, the cluster labels obtained can be mapped to the dataset. This converts the problem into a multi-class classification problem and models can be built over it.

7.2 Binary Classification Models

7.2.1 Model Building

- The dataset was split into train and test sets by considering the train test ratio as 70:30.
- These two sets were then scaled and used for modelling.
- As the dataset is balanced, the evaluation metrics considered are ROC AUC and Accuracy scores.
- Various supervised classification algorithms like Logistic Regression, Decision Tree, Random Forest, Adaptive Boosting, Gradient Boosting and Support Vector Machine were used to build the models.
- Also, Principal Component Analysis was applied to the dataset and models were built upon the transformed data.
- To enhance the performance of the models, feature selection techniques were applied and hyperparameter tuning for various models has been done using Grid Search CV or Randomized Search CV
- Finally, the evaluation of best performing models was done using 5-fold cross validation.

	ROC AUC		ACCURACY	
MODELS	**Train**	**Test**	**Train**	**Test**
Logistic Regression (Base Model)	70.65	69.75	65.61	64.31
Decision Tree	65.82	64.98	62.89	62.42
Random Forest	70.13	69.29	64.71	64.25
Support Vector Machine	69.99	68.99	64.49	63.31
RFE + Logistic Regression	69.19	68.43	64.45	62.95
RFE + Decision Tree	65.68	64.98	62.88	62.15
RFE + Random Forest	70.68	69.70	64.84	64.56
RFE + Support Vector Machine	78.10	70.50	70.11	64.82

Table 9: Initial Models

- Further Boosting algorithms like AdaBoost and Gradient Boost were implemented to increase the accuracy.

	ROC AUC		ACCURACY	
MODELS	**Train**	**Test**	**Train**	**Test**
Boosted RFE + DT	72.94	69.88	66.79	64.27
Boosted RFE + RF	74.49	71.84	67.87	65.79
Gradient Boost	71.67	70.23	65.88	64.99
Boosted RFE + Support Vector Machine	68.03	67.14	60.48	59.71

Table 10: Boosted Models

- Some of the above models were tuned in order to improve their performance.

	ROC AUC		ACCURACY	
MODELS	**Train**	**Test**	**Train**	**Test**
RFE + Decision Tree (Tuned)	70.76	68.28	65.11	63.09
RFE + Random Forest (Tuned)	73.07	69.99	67.20	64.57
Tuned Decision Tree (20)	70.74	68.31	65.05	63.09
Tuned Random Forest (20)	72.25	69.55	66.28	64.23
Boosted RFE + DT (Tuned)	73.31	70.00	66.93	64.32
Boosted RFE + RF (Tuned)	75.36	72.01	68.65	65.87
Gradient Boost (Tuned)	75.05	71.87	68.35	65.98

Table 11: Tuned Models

- After tuning, there is an increase in the roc auc and accuracy scores by 1 or 2%.
- The ensemble models are performing better among all the models built.
- The highest roc auc score and accuracy scores are given by Boosted RFE + DT (Tuned), Boosted RFE + RF (Tuned) and Gradient Boost (Tuned).
- Further to enhance the performance, these models were combined.
- Considered the following models for Stacking,
    - Stacked Model 1: Boosted RFE + DT (Tuned) / Boosted RFE + RF (Tuned) / Gradient Boost (Tuned)
    - Stacked Model 2: Boosted RFE + RF (Tuned) / Gradient Boost (Tuned)

	ROC AUC		ACCURACY	
MODELS	**Train**	**Test**	**Train**	**Test**
Boosted RFE + DT (Tuned)/ Boosted RFE + RF (Tuned)/ Gradient Boost (Tuned)	75.14	71.90	68.33	65.92
Boosted RFE + RF (Tuned)/ Gradient Boost (Tuned)	75.10	71.90	68.33	66.01

Table 12: Stacked Models

- The models were built after applying dimensionality reduction technique (PCA) but did not improve the performance of the models.

	ROC AUC		ACCURACY	
MODELS	**Train**	**Test**	**Train**	**Test**
PCA + Logistic Regression	69.58	68.91	64.59	63.55
PCA + Decision Tree	63.55	62.91	59.98	59.68
PCA + Random Forest	67.95	66.72	63.21	62.20

Table 13: PCA Models

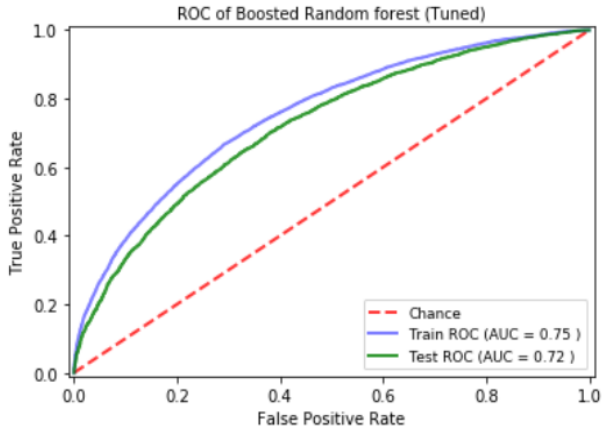

Figure 43: Receiver Operating Characteristics of Boosted Random Forest (Tuned)

Classification Report:

	precision	recall	f1-score	support
0	0.66	0.67	0.67	6072
1	0.65	0.65	0.65	5822
accuracy			0.66	11894
macro avg	0.66	0.66	0.66	11894
weighted avg	0.66	0.66	0.66	11894

Figure 44: Classification Report of Boosted Random Forest (Tuned)

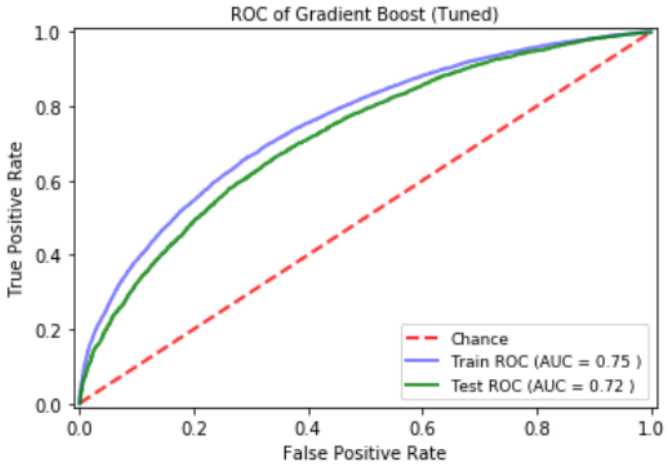

Figure 45: Receiver Operating Characteristics of Gradient Boost (Tuned)

Classification Report:

	precision	recall	f1-score	support
0	0.66	0.67	0.67	6072
1	0.65	0.65	0.65	5822
accuracy			0.66	11894
macro avg	0.66	0.66	0.66	11894
weighted avg	0.66	0.66	0.66	11894

Figure 46: Classification Report of Gradient Boost (Tuned)

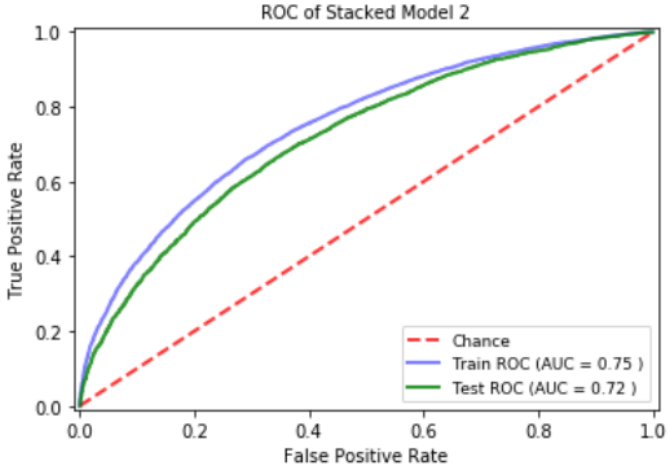

Figure 47: Receiver Operating Characteristics of Stacked Model 2

```
Classification Report-Test:
              precision    recall  f1-score   support

           0       0.67      0.67      0.67      6144
           1       0.65      0.65      0.65      5750

    accuracy                           0.66     11894
   macro avg       0.66      0.66      0.66     11894
weighted avg       0.66      0.66      0.66     11894
```

Figure 48: Classification Report of Stacked Model 2

7.2.2 Model Evaluation:

From the above analysis of scores for various models, the final models selected for evaluation are:

1. Boosted Random Forest (Tuned)
2. Gradient Boost (Tuned)
3. Stacked Model 2 (Boosted Random Forest (Tuned) + Gradient Boost (Tuned))

These models were evaluated by using 5-fold cross validation, with roc_auc as a scoring function. The train and test scores for each model were recorded to check for any kind of overfitting or underfitting. Finally, the mean of the test scores were calculated to decide which model fits the best.

MODELS	MEAN ROC AUC SCORE	MEAN ACCURACY
Boosted RFE + RF (Tuned)	73	66.77
Gradient Boost (Tuned)	72.76	66.72
Boosted RFE + RF (Tuned)/ Gradient Boost (Tuned)	72.81	66.71

Table 14: Scores of Evaluated Models

- The accuracy value of Boosted RFE + RF (Tuned) model indicates that 66.77% of the data is correctly predicted and the roc auc score indicates that there is 73% chance that the model is able to distinguish between the popular and unpopular articles. Hence, it is a good model for the Mashable dataset among the other models.

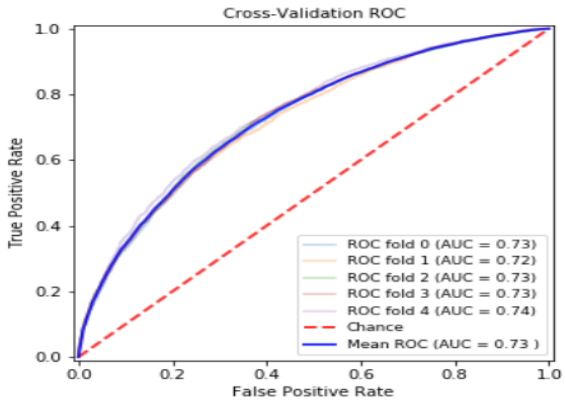

Figure 49: ROC for 5-fold CV of Boosted RFE + RF (Tuned) model

7.3 Multi-Class Classification Models

7.3.1 PCA and Clustering

- The original dataset has 61 attributes including target (shares) out of which 60 were numeric and 1 attribute (url) was of string data type.
- The 60 scaled attributes were used for Principal Component Analysis (PCA) which would be further used for Unsupervised approach (KMeans).

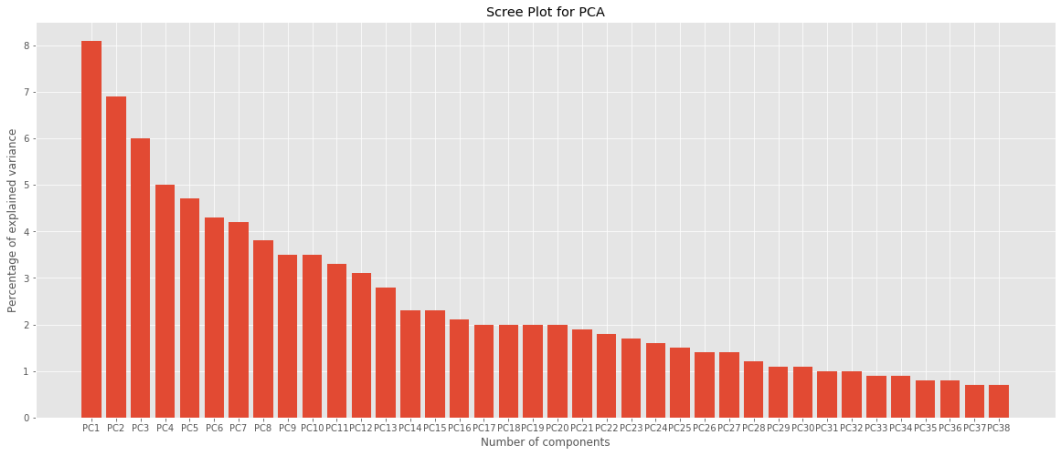

Figure 50: Number of PC's

7.3.2 KMeans Algorithm

- The scaled Principal components totally having a variance of 95% of the original data was fed into KMeans algorithm.

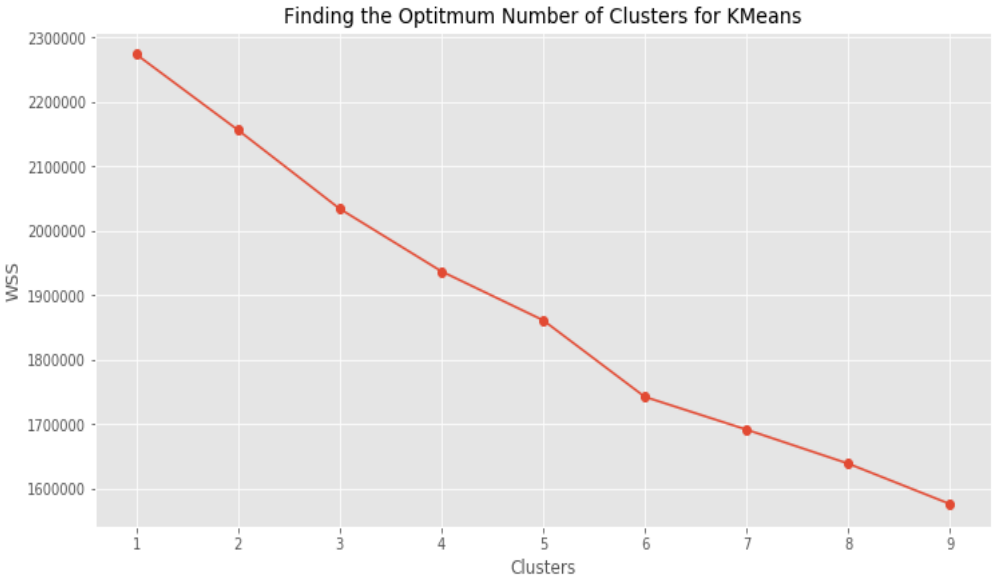

Figure 51: Finding the Optimum number of Clusters for KMeans algorithm

- From the above plot the number of clusters chosen was 5. This value of 'k' was used as input to the KMeans algorithm. The Cluster labels retrieved were used as a target variable for multi-classification.

7.3.3 Cluster Analysis:

- The clusters labels derived from KMeans algorithms have strong correlation with the following attributes:

Cluster 0:

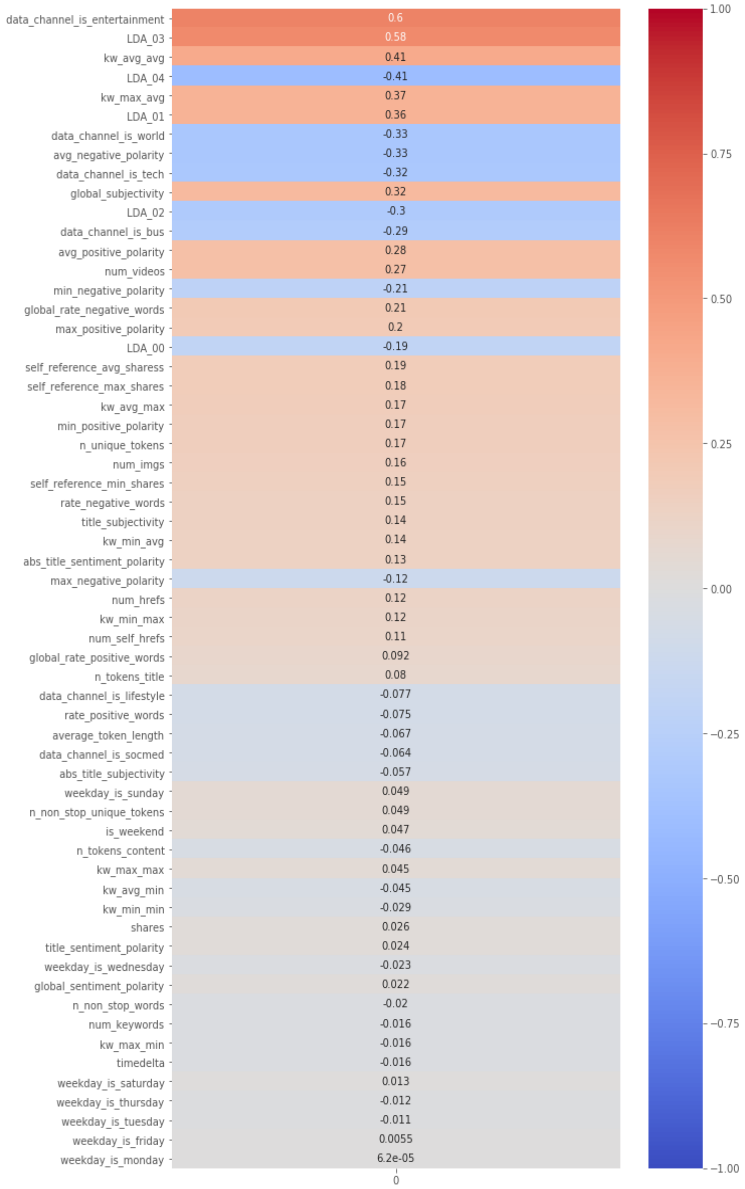

Figure 52: Correlation between attributes for Cluster 0

Cluster 0 has a strong relation with Entertainment data channel, LDA_03

Cluster 1:

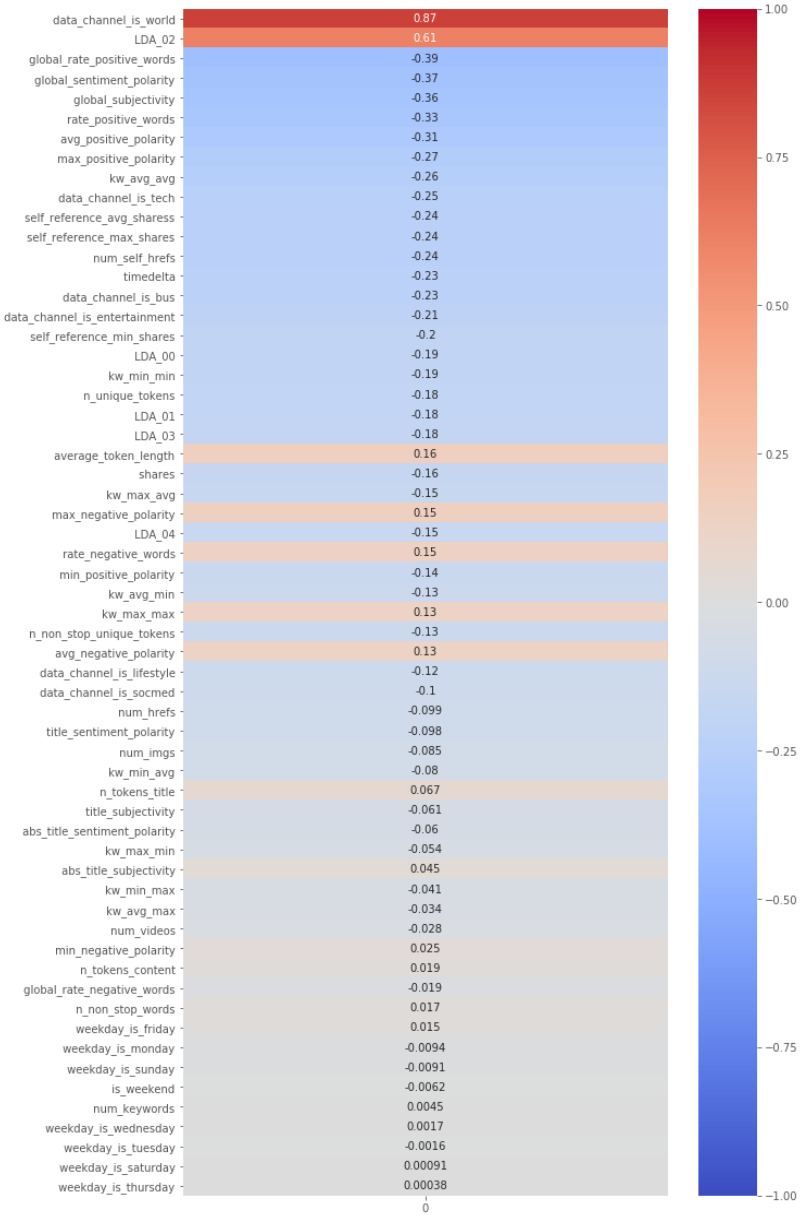

Figure 53: Correlation between attributes for Cluster 1

Cluster 1 has a strong relation with World data channel, LDA_02

Cluster 2:

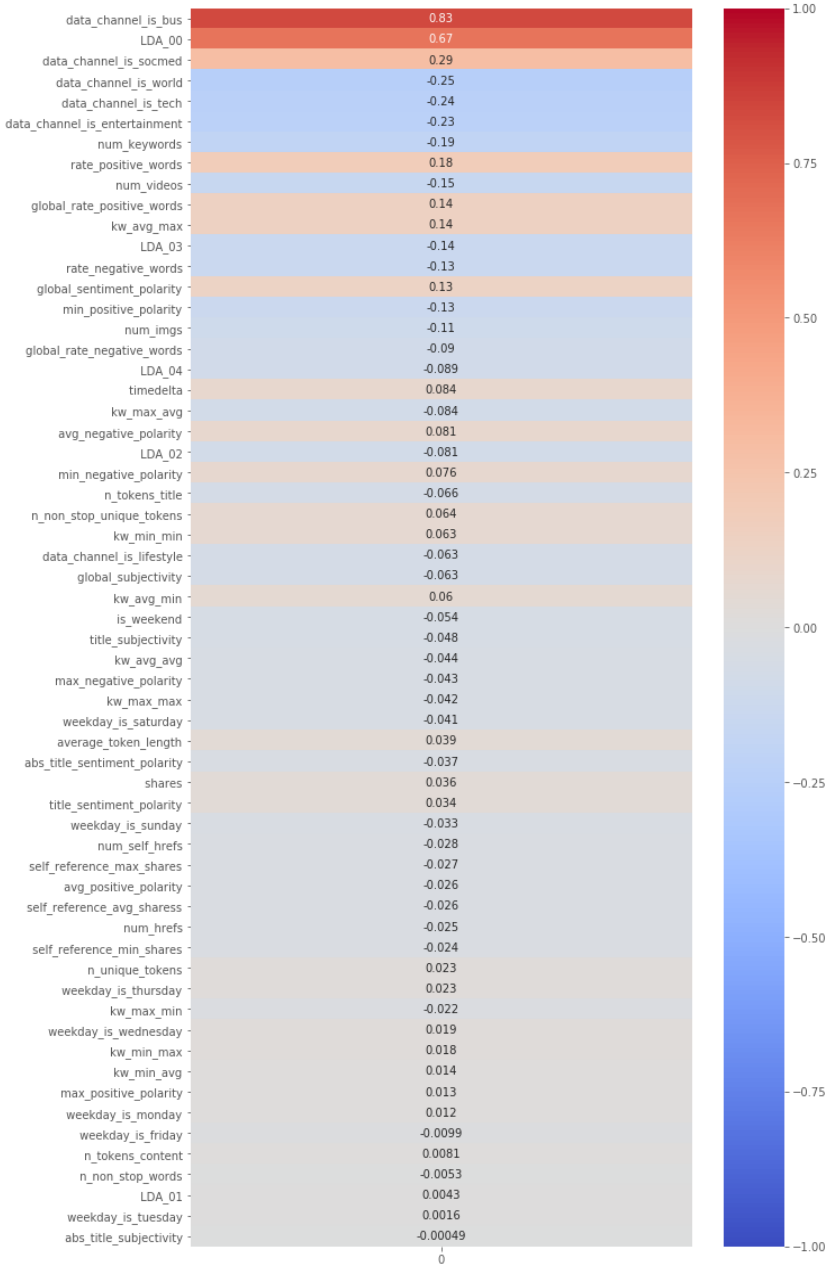

Figure 54: Correlation between attributes for Cluster 2

Cluster 2 has a strong relation with Business data channel, LDA_00

Cluster 3:

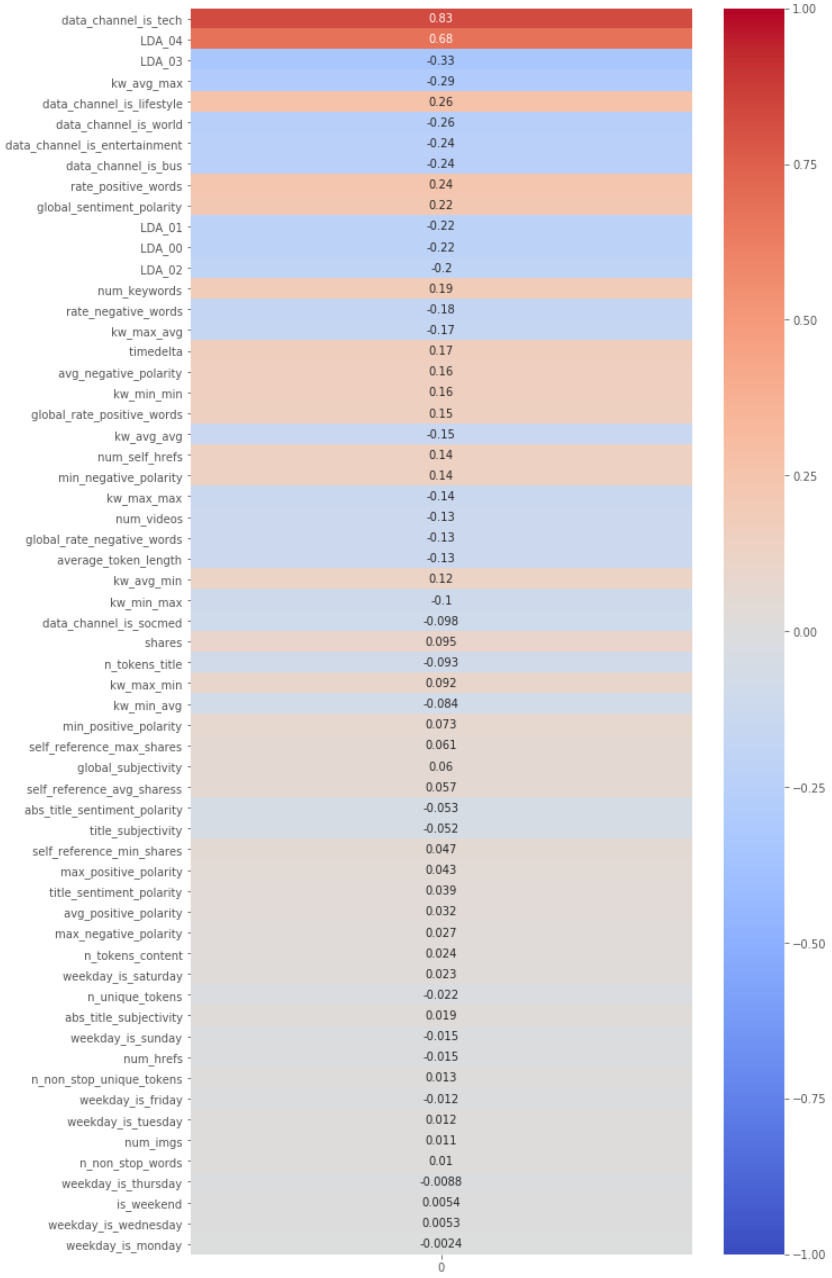

Figure 55: Correlation between attributes for Cluster 3

Cluster 3 has a strong relation with Tech data channel, LDA_02

Cluster 4:

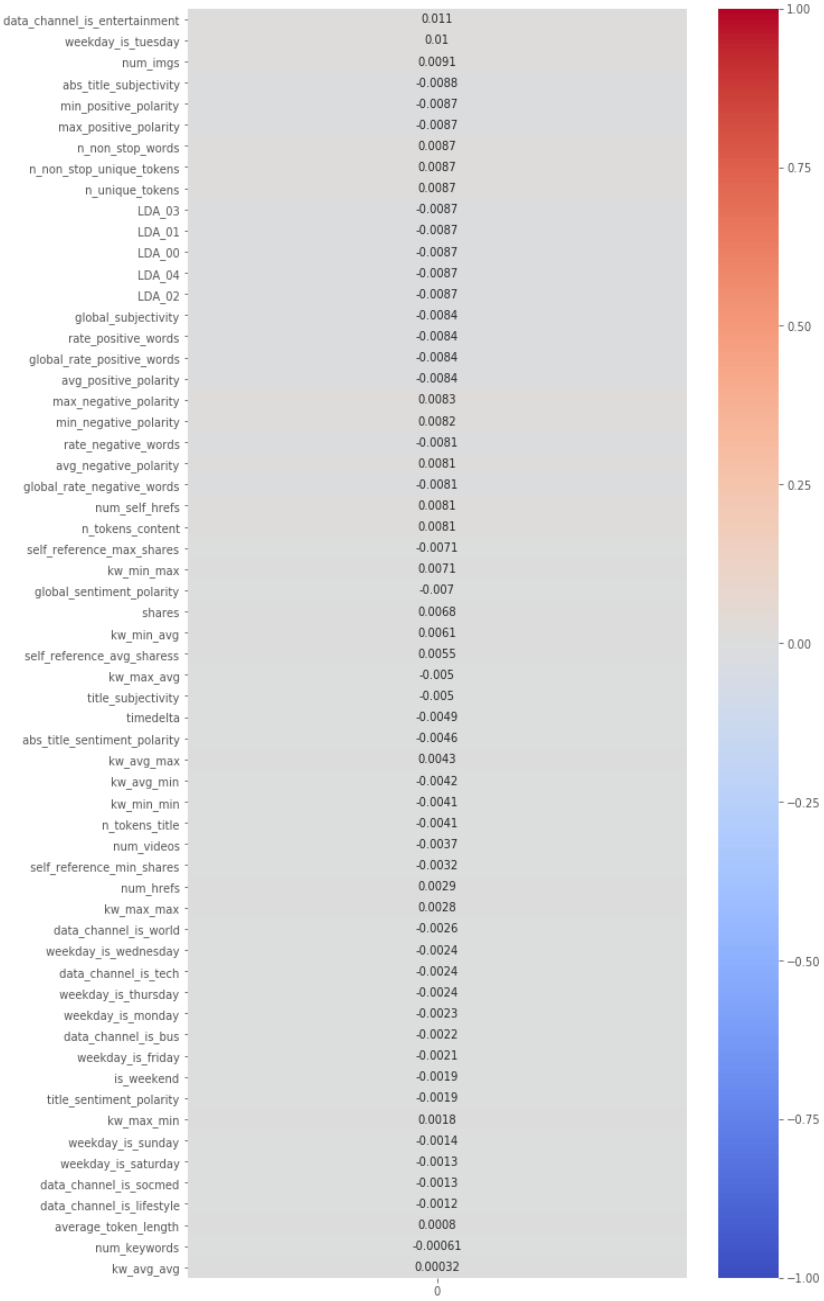

Figure 56: Correlation between attributes for Cluster 3

Cluster 4 has no relation with a known data channel and LDA

7.3.4 Models Built

The dataset was split into train and test sets by considering the train test ratio as 70:30 and considering random_state as 0.

- Check for imbalance
- These two sets were then scaled and used for modelling.
- The evaluation metrics considered are F1-weighted score, Classification Report, Confusion Matrix and Kappa Cohen score.

APPROACH	FEATURE ENGINEERING	EVALUATION METRICS	BASE MODEL		BAGGING		BOOSTING			
							ADA-BOOST		GRADIENT-BOOST	XG-BOOST
		MODEL	Decision Tree	Random Forest	Decision Tree	Random Forest	Decision Tree	Random Forest	Decision Tree	Decision Tree
KMEANS + SCALED DATASET (SMOTE) (60+1 features)	ALL FEATURES	TRAIN SCORE	0.96	0.98	0.97	0.98	0.98	0.98	0.97	0.99
		TEST SCORE	0.93	0.96	0.96	0.96	0.96	0.97	0.96	0.97
		KAPPA COHEN TRAIN SCORE	0.95	0.97	0.97	0.97	0.97	0.98	0.97	0.98
		KAPPA COHEN TEST SCORE	0.89	0.93	0.93	0.93	0.93	0.95	0.94	0.96
PCA + KMEANS + SCALED DATASET (60+1 features)	ALL FEATURES	TRAIN SCORE	0.95	0.96	0.96	0.96	0.96	0.97	0.97	0.98
		TEST SCORE	0.94	0.96	0.95	0.96	0.96	0.97	0.97	0.97
		KAPPA COHEN TRAIN SCORE	0.93	0.95	0.94	0.95	0.95	0.96	0.96	0.97
		KAPPA COHEN TEST SCORE	0.92	0.95	0.94	0.95	0.94	0.95	0.95	0.97
PCA + KMEANS + SCALED DATASET (60+1 features)	ALL FEATURES	CROSS VALIDATION SCORE	0.95	0.96	0.96	0.96	0.96	0.97	0.97	0.98
PCA + KMEANS + PCA COMPONENTS (38+1 features)	PCA	TRAIN SCORE	0.93	0.96	0.96	0.96	0.97	0.97	0.96	0.98
		TEST SCORE	0.9	0.96	0.95	0.96	0.96	0.96	0.96	0.97
		KAPPA COHEN TRAIN SCORE	0.90	0.95	0.94	0.95	0.95	0.96	0.95	0.97
		KAPPA COHEN TEST SCORE	0.87	0.94	0.93	0.95	0.95	0.95	0.95	0.97
PCA + KMEANS + SCALED DATASET (17+1 features)	FEATURE IMPORTANCE	TRAIN SCORE	0.95	0.96	0.96	0.96	0.95	0.96	0.96	0.96
		TEST SCORE	0.94	0.96	0.95	0.95	0.95	0.96	0.96	0.96
		KAPPA COHEN TRAIN SCORE	0.93	0.95	0.94	0.95	0.94	0.94	0.95	0.95
		KAPPA COHEN TEST SCORE	0.93	0.94	0.93	0.94	0.93	0.94	0.94	0.95
PCA + KMEANS + SCALED DATASET (30+1 features)	RFE	TRAIN SCORE	0.95	0.96	0.96	0.96	0.96	0.97	0.97	0.98
		TEST SCORE	0.94	0.96	0.95	0.96	0.96	0.97	0.96	0.97
		KAPPA COHEN TRAIN SCORE	0.93	0.95	0.94	0.95	0.95	0.96	0.96	0.97
		KAPPA COHEN TEST SCORE	0.92	0.95	0.94	0.95	0.95	0.95	0.95	0.96
PCA + KMEANS + SCALLD DATASET (13 features)	RFE CV	TRAIN SCORE	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.96
		TEST SCORE	0.93	0.95	0.94	0.95	0.95	0.95	0.95	0.96
		KAPPA COHEN TRAIN SCORE	0.92	0.94	0.93	0.94	0.93	0.94	0.94	0.95
		KAPPA COHEN TEST SCORE	0.91	0.94	0.92	0.94	0.93	0.93	0.94	0.94
PCA + KMEANS + SCALED DATASET (3+1 features)	CORRELATION MATRIX ANALYSIS	TRAIN SCORE	0.77	0.79	0.78	0.79	0.76	0.76	0.76	0.78
		TEST SCORE	0.76	0.78	0.78	0.78	0.75	0.71	0.76	0.78
		KAPPA COHEN TRAIN SCORE	0.68	0.71	0.71	0.71	0.67	0.68	0.68	0.70
		KAPPA COHEN TEST SCORE	0.67	0.71	0.70	0.70	0.66	0.61	0.68	0.70
PCA + KMEANS + SCALED DATASET (17+1 features)	SELECT K BEST (CHI SQUARE)	TRAIN SCORE	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96
		TEST SCORE	0.94	0.95	0.95	0.95	0.95	0.95	0.96	0.96
		KAPPA COHEN TRAIN SCORE	0.93	0.94	0.93	0.94	0.93	0.94	0.94	0.95
		KAPPA COHEN TEST SCORE	0.92	0.94	0.93	0.93	0.93	0.94	0.94	0.94
PCA + KMEANS + SCALED DATASET (17+1 features)	SELECT K BEST (MUTUAL INFORMATION)	TRAIN SCORE	0.94	0.96	0.96	0.96	0.96	0.96	0.96	0.96
		TEST SCORE	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.96
		KAPPA COHEN TRAIN SCORE	0.92	0.94	0.94	0.94	0.94	0.94	0.94	0.95
		KAPPA COHEN TEST SCORE	0.91	0.94	0.93	0.94	0.94	0.94	0.94	0.95

Figure 57: Modelling Results

From the figure, it is known that 10 approaches have been adopted for multi-classification modelling using the unsupervised approach (KMeans).

1. Multi Classification using Unsupervised approach (KMeans) with Scaled 60 (Independent) and 1 (Dependent) features:
    a. The 'url' attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled.
    b. KMeans algorithm was applied on the dataset and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (60 attributes) was attached to the cluster labels (target) and multi-classification was applied.

d. But before that the target class labels (cluster labels) were imbalanced. So, they were addressed using SMOTE analysis.

Figure 58: Class Imbalance

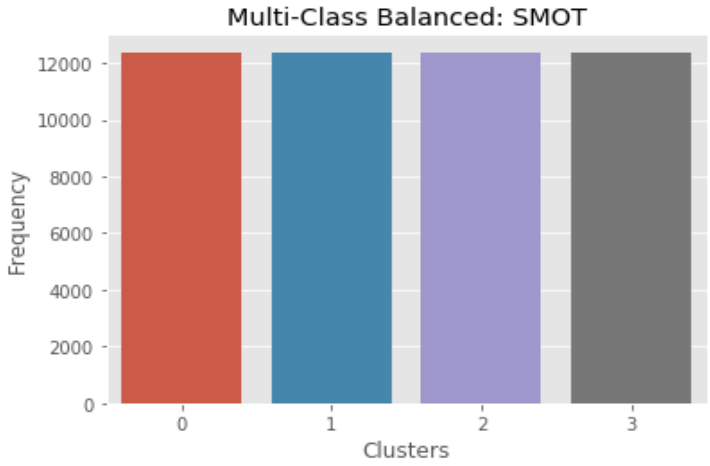

Figure 59: Balanced Class Labels using SMOT

e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
g. Further, bagging and boosting were applied on the base models.
h. The preferred model chosen was ada-boost for random forest which had a f1-weighted train score of 0.98, test score of 0.97 and kappa cohen train score of 0.98 and test score of 0.95

2. Multi Classification using Unsupervised approach (PCA+KMeans) with Scaled 60 (Independent) and 1 (Dependent) features:
    a. The 'url' attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC's) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (60 attributes) was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

Figure 60: Balanced Class Labels

    e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
    f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
    g. Further, bagging and boosting were applied on the base models.
    h. The preferred model chosen was ada-boost for random forest which had a f1-weighted train score of 0.97, test score of 0.97 and kappa cohen train score of 0.96 and test score of 0.95

3. Multi Classification using Unsupervised approach (PCA+KMeans) with Scaled 60 (Independent) and 1 (Dependent) features:
    a. The 'url' attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC's) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (60 attributes) was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

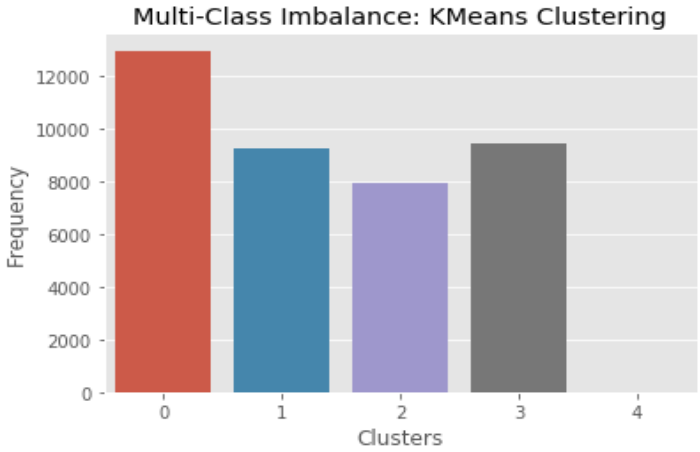

Figure 61: Balanced Class Labels

e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
g. Further, bagging and boosting were applied on the base models.
h. The preferred model chosen was xg-boost for random forest which had a f1-weighted cross validation score is 0.98

4. Multi Classification using Unsupervised approach (KMeans) with Scaled 38 (Independent) Principal Components (PC) and 1 (Dependent) features:
    a. The 'url' attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC's) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The PC (38 attributes) was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

Figure 62: Balanced Class Labels

e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
g. Further, bagging and boosting were applied on the base models.
h. The **BEST MODEL** chosen out of all the given approaches was xg-boost for decision tree which had a f1-weighted train score of 0.98, test score of 0.97 and kappa cohen train score of 0.97 and test score of 0.97

5. Multi Classification using Unsupervised approach (KMeans) with Scaled 17 (Independent) and 1 (Dependent) features:
    a. The 'url' attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC's) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (17 attributes) based on Feature Importance was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

Figure 63: Balanced Class Labels

e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
g. Further, bagging and boosting were applied on the base models.
h. The preferred model chosen was xg-boost for decision tree which had a f1-weighted train score of 0.96, test score of 0.96 and kappa cohen train score of 0.95 and test score of 0.95

6. Multi Classification using Unsupervised approach (KMeans) with Scaled 30 (Independent) and 1 (Dependent) features:
    a. The ‘url’ attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC’s) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (13 attributes) based on Recursive Feature Elimination with Cross Validation (RFE) was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

Figure 64: Balanced Class Labels

    e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
    f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
    g. Further, bagging and boosting were applied on the base models.
    h. The preferred model chosen was ada-boost for random forest which had a f1-weighted train score of 0.97, test score of 0.97 and kappa cohen train score of 0.96 and test score of 0.95

7. Multi Classification using Unsupervised approach (KMeans) with Scaled 13 (Independent) and 1 (Dependent) features:
    a. The ‘url’ attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC’s) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (13 attributes) based on Recursive Feature Elimination with cross validation (RFECV) was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

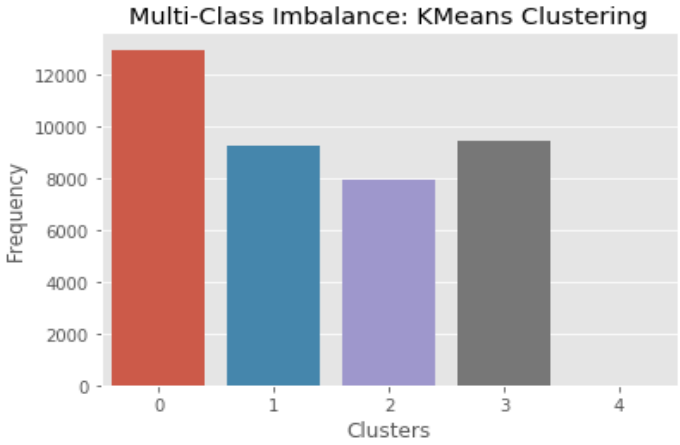

Figure 65: Balanced Class Labels

e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
g. Further, bagging and boosting were applied on the base models.
h. The preferred model chosen was xg-boost for decision tree which had a f1-weighted train score of 0.96, test score of 0.96 and kappa cohen train score of 0.95 and test score of 0.94

8. Multi Classification using Unsupervised approach (KMeans) with Scaled 3 (Independent) and 1 (Dependent) features:
    a. The 'url' attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC's) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The dataset (3 attributes) based on Correlation matrix analysis was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

Figure 66: Balanced Class Labels

e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
g. Further, bagging and boosting were applied on the base models.
h. The preferred model chosen was bagging for decision tree which had a f1-weighted train score of 0.78, test score of 0.78 and kappa cohen train score of 0.71 and test score of 0.70

9. Multi Classification using Unsupervised approach (KMeans) with Scaled 17 (Independent) and 1 (Dependent) features:
    a. The 'url' attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC's) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (17 attributes) based on SelectKBest (chi-square) was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

Figure 67: Balanced Class Labels

e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
g. Further, bagging and boosting were applied on the base models.
h. The preferred model chosen was gradient-boost for decision tree which had a f1-weighted train score of 0.96, test score of 0.96 and kappa cohen train score of 0.94 and test score of 0.94

10. Multi Classification using Unsupervised approach (KMeans) with Scaled 17 (Independent) and 1 (Dependent) features:
    a. The ‘url’ attribute of string data-type was removed, so we had a dataset of 60 attributes which was scaled and Principal Component Analysis (PCA) was performed.
    b. KMeans algorithm was applied on the Principal Components (PC’s) and 5 clusters were obtained. The cluster labels were used as the target for Supervised Learning.
    c. The scaled dataset (17 attributes) based on SelectKBest (mutual-information) was attached to the cluster labels (target) and multi-classification was applied.
    d. The target class labels (cluster labels) were balanced.

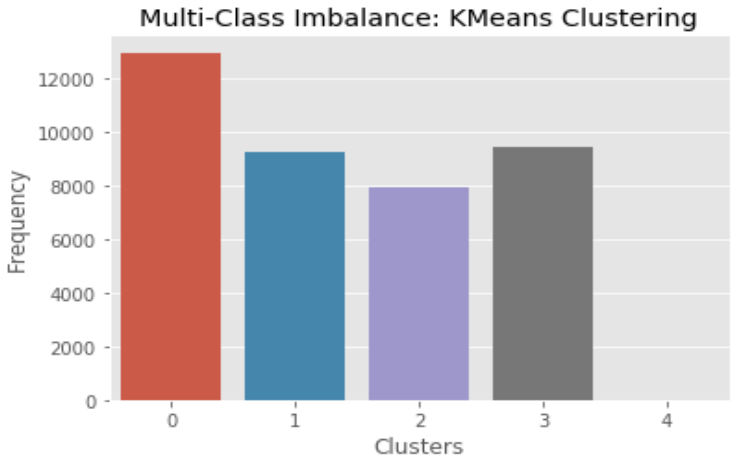

Figure 68: Balanced Class Labels

    e. Randomized Search Cross Validation (RSCV) was applied on the Train data and the optimal parameters were chosen for the decision tree as the base model which was further hyper-parameter tuned.
    f. Hyper-parameter tuned decision trees and random forest were used as base models for Multi-classification.
    g. Further, bagging and boosting were applied on the base models.
    h. The preferred model chosen was xg-boost for decision tree which had a f1-weighted train score of 0.96, test score of 0.96 and kappa cohen train score of 0.95 and test score of 0.95

7.3.5 Results

Multi Classification using Unsupervised approach (KMeans) with Scaled 38 (Independent) Principal Components (PC) and 1 (Dependent) features gave the best result using xg-boost for decision tree.

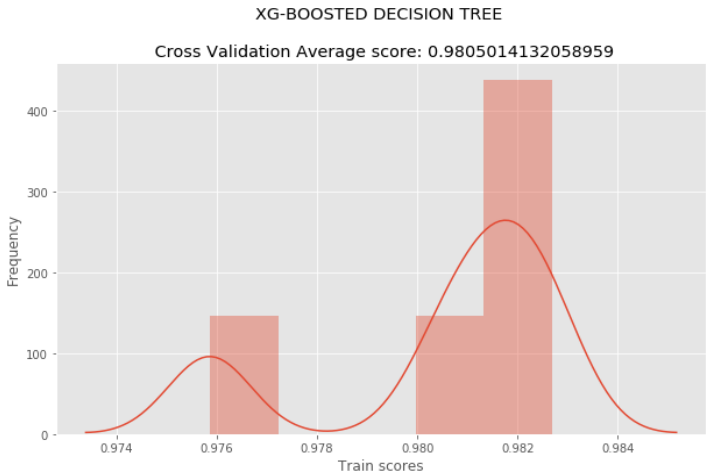

Figure 69: Randomised Search Cross Validation (RSCV) on Train data

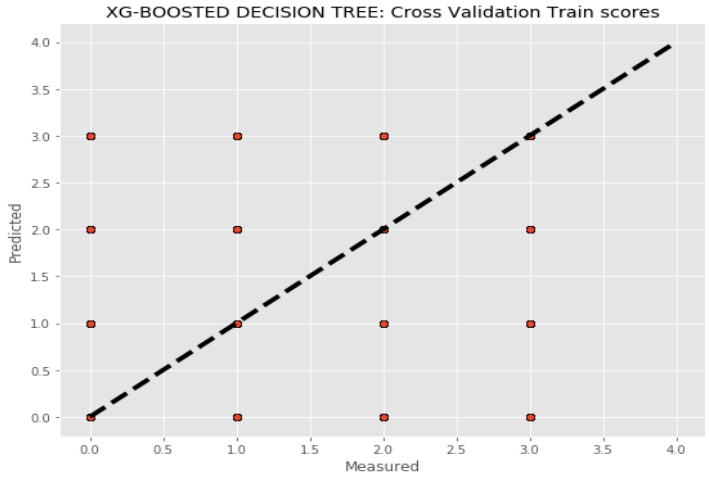

Figure 70: Cross Validation Train Scores for XG-Boosted Decision Tree

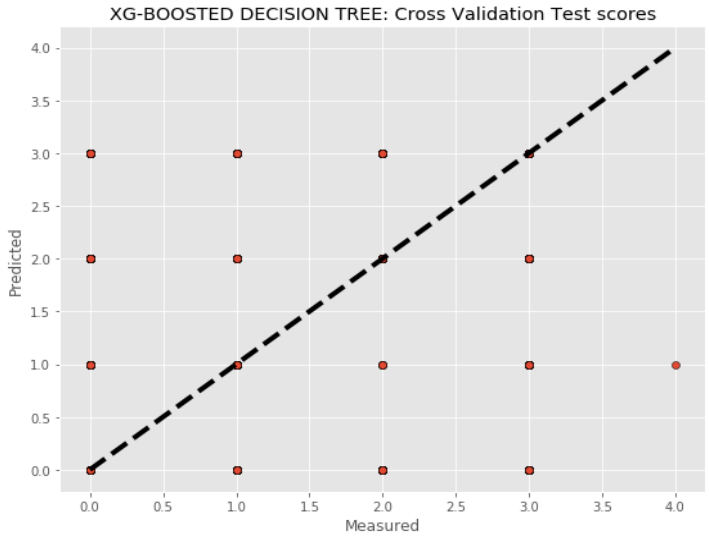

Figure 71: Cross Validation Train Scores for XG-Boosted Decision Tree

	precision	recall	f1-score	support
0	0.98	0.99	0.98	9099
1	0.99	0.98	0.98	6524
2	0.98	0.98	0.98	5558
3	0.98	0.97	0.98	6569
accuracy			0.98	27750
macro avg	0.98	0.98	0.98	27750
weighted avg	0.98	0.98	0.98	27750

Figure 72: XG-Boosted Decision Tree Classification Report for Train data

	precision	recall	f1-score	support
0	0.97	0.98	0.98	3879
1	0.98	0.97	0.98	2742
2	0.97	0.97	0.97	2377
3	0.98	0.97	0.97	2895
4	0.00	0.00	0.00	1
accuracy			0.97	11894
macro avg	0.78	0.78	0.78	11894
weighted avg	0.97	0.97	0.97	11894

Figure 73: XG-Boosted Decision Tree Classification Report for Test data

```
[[8982   26   56   35]
 [  70 6387   34   33]
 [  68   19 5422   49]
 [  91   50   44 6384]]
```

Figure 74: XG-Boosted Decision Tree Confusion Matrix for Train data

```
[[3812   19   27   21    0]
 [  45 2660   19   18    0]
 [  33    6 2314   24    0]
 [  46   22   24 2803    0]
 [   0    1    0    0    0]]
```

Figure 75: XG-Boosted Decision Tree Confusion Matrix for Test data

- Kappa Cohen Train score: 0.9720251172641339
- Kappa Cohen Test score: 0.9653882359177652

The f1-weighted train score of 0.98, test score of 0.97 and kappa cohen train score of 0.97 and test score of 0.97 from XG Boosted Decision Tree seemed to be the most preferred for Multi-Classification of News Articles based on category (LDA, Data Channel).

7.4 Conclusion

Binary Classification - Classification based on Shares

The Boosted RFE + RF (Tuned) model with an accuracy of 66.77% and the ROC AUC score of 73% is the best model for classifying the articles published by Mashable as popular and unpopular.

Multi Classification - Classification based on Unsupervised Approach

The f1-weighted train score of 0.98, test score of 0.97 and kappa cohen train score of 0.97 and test score of 0.97 from XG Boosted Decision Tree seemed to be the most preferred for Multi-Classification of News Articles based on category (LDA, Data Channel).

Recommendation and Improvements

Recommendations:

For an article to be Popular:

- The title should have 7 -15 words
- The articles should be short (382-2591 words)
- Less number of images and videos (0-2)
- Less number of links (3-5)
- Include articles that have positive sentiments that can relate to people
- Increase the Release of the articles on weekends, as they are popular
- Can reduce the number of news articles on weekdays
- Increase the number of articles related to LDA_04 (Technology data channel) and reduce for LDA_02(World data channel)
- The news article content should be factual for world data channel
- Articles related to Social media / Lifestyle data channel are popular, so increase the number of articles released

Improvements:

Addition of more data based on the following:

- Find data related to publication time (date, month, year, timestamp)
- Find the keywords for each article so that Sentiment Analysis, NLP can be applied on the data to further improve insights on online news articles.
- Make known the names of the LDA topics as well as their characteristics to better understand news article categories.
- Find the polarity (positive and negative) of keywords for better insights on the news articles and enhance popularity.