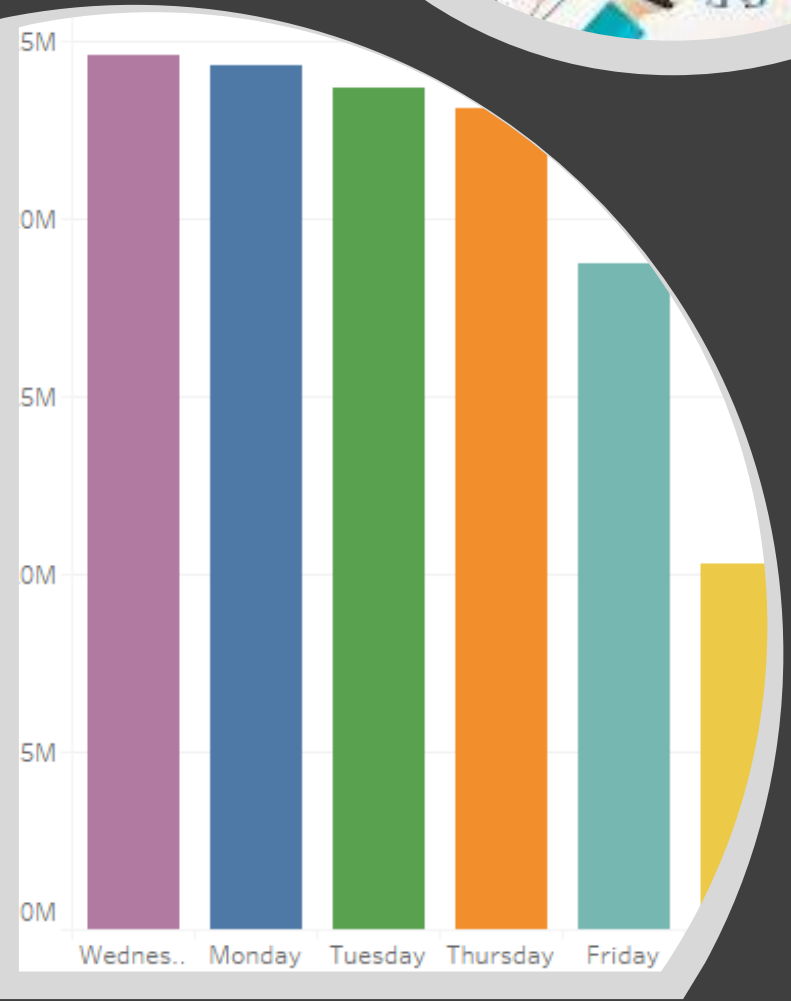
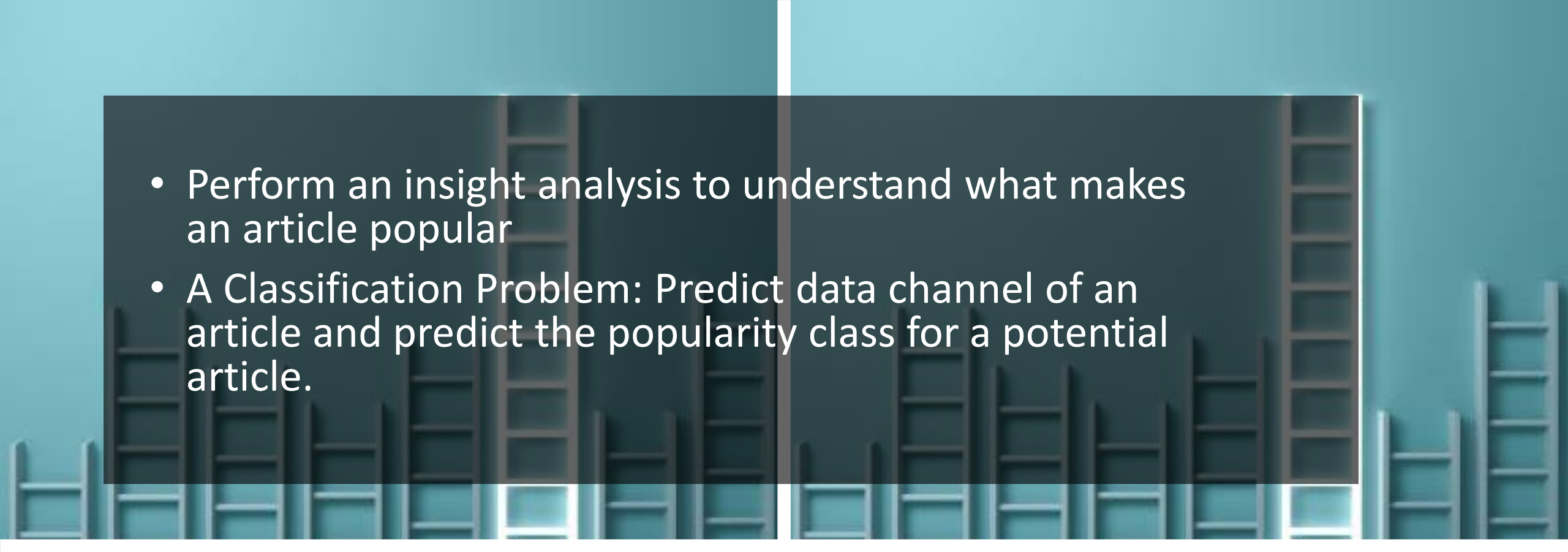


Exploratory Analysis of Online News Popularity



- Perform an insight analysis to understand what makes an article popular
- A Classification Problem: Predict data channel of an article and predict the popularity class for a potential article.

Our Goal

In the digital era, people like reading, writing, and sharing articles online, but what makes some articles very popular compared to others in spite of quality work is the question we like to address as part of this project.

Also, we will address the article dataset as a classification problem: predict the data channel and the popularity of an article

Dataset

- This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years
- Records: 39644
- Attributes: 61

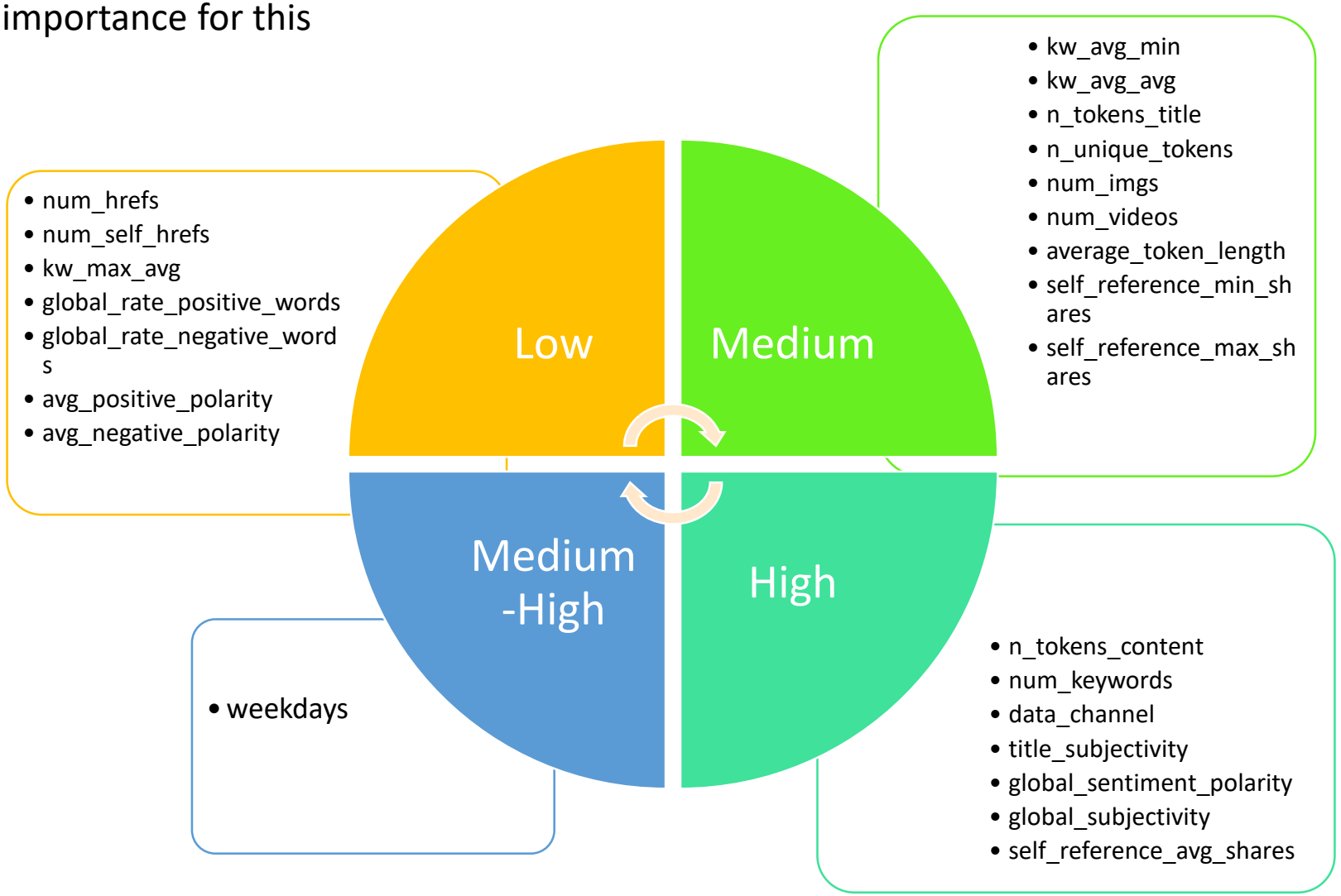
url	timedelta	n_tokens	n_tokens	n_unique	n_non_st	n_non_st	num_hre	num_self	num_img	num_vide
http://ma	731	12	219	0.663594	1	0.815385	4	2	1	0
http://ma	731	9	255	0.604743	1	0.791946	3	1	1	0
http://ma	731	9	211	0.57513	1	0.663866	3	1	1	0
http://ma	731	9	531	0.503788	1	0.665635	9	0	1	0
http://ma	731	13	1072	0.415646	1	0.54089	19	19	20	0
http://ma	731	10	370	0.559889	1	0.698198	2	2	0	0
http://ma	731	8	960	0.418163	1	0.549834	21	20	20	0
http://ma	731	12	989	0.433574	1	0.572108	20	20	20	0
http://ma	731	11	97	0.670103	1	0.836735	2	0	0	0
http://ma	731	10	231	0.636364	1	0.797101	4	1	1	1
http://ma	731	9	1248	0.49005	1	0.731638	11	0	1	0
http://ma	731	10	187	0.666667	1	0.8	7	0	1	0
http://ma	731	9	274	0.609195	1	0.707602	18	2	11	0
http://ma	731	9	285	0.744186	1	0.84153	4	2	0	21

Attributes in Dataset:

Category	Variables
Based on words	n_tokens_title,n_tokens_content, num_keyword,n_unique_tokens: n_non_stop_words, n_non_stop_unique_tokens
Reference	num_href, num_self_href, self_reference
Visuals	Num_imgs, num_videos
Binary output	Data_channel, weekday
Text mining/NLP	Kw_words, LDA topics, global subjectivity, global sentiment polarity, positive and negative words(global and rate), positive_polarity, negative_polarity, title subjectivity, title sentiment polarity
Target	Shares

Our Hypothesis – A Subjective Analysis

We looked at each variable and did a philosophical analysis about their meaning and importance for this problem.



Data Preprocessing

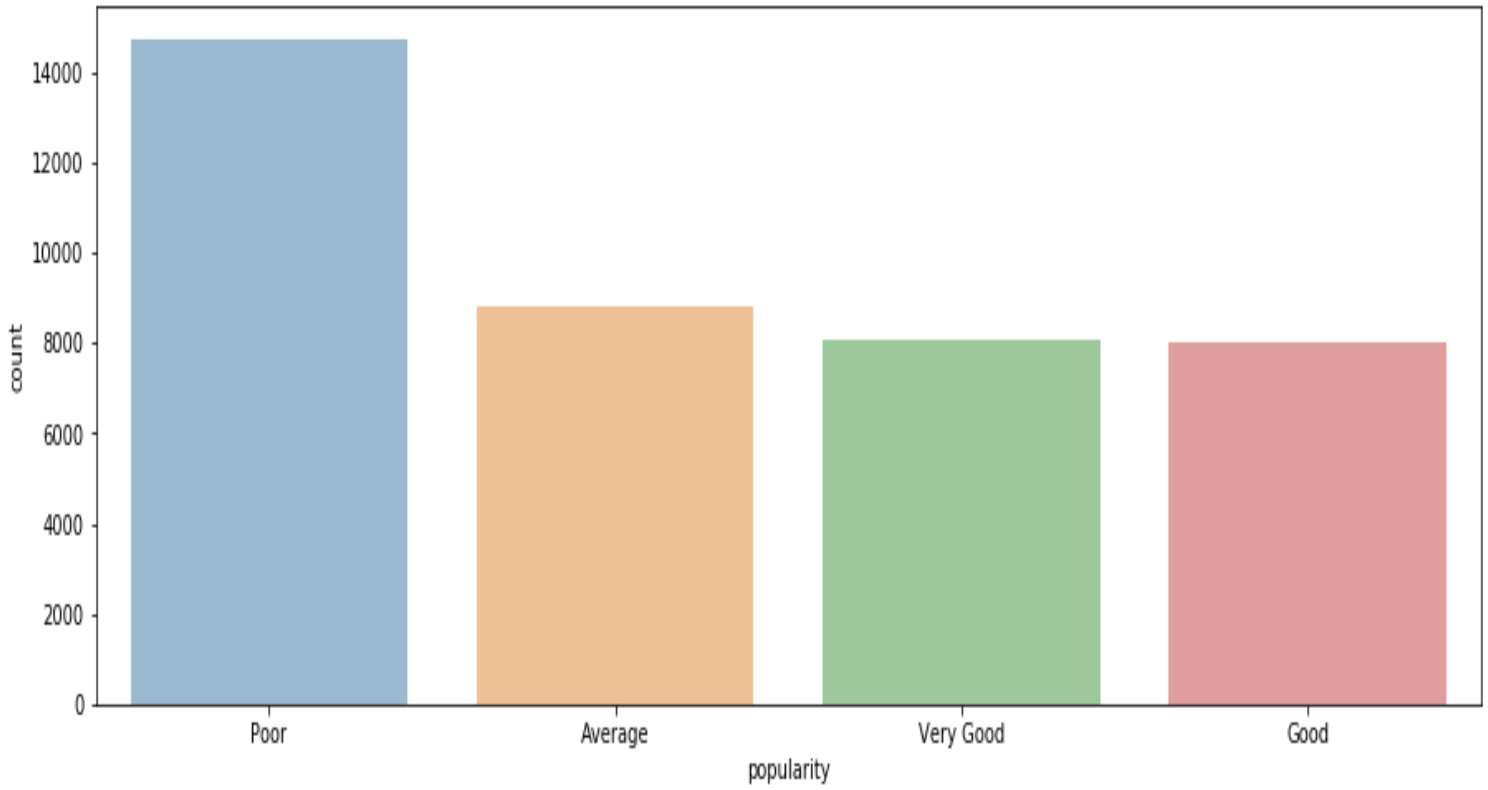
- Seven weekday columns which had binary output was merged into one column
- Six data channel colums which had binary output was merged into one column
- Dropped 2 columns- url and timedelta
- Noise from 2 columns was removed (n_stop_words and n_tokens_length)

- Very Good = Top 80%
- Good = Top 60% -Top 80%
- Average = Top 40% -Top 60%
- Poor = Below 40%

	class	shares	No. of records
1	Poor	Less than 1200	14346
2	Average	Between 1200 and 1800	8585
3	Good	Between 1800 and 3400	7785
4	Very Good	Greater than 3400	7746

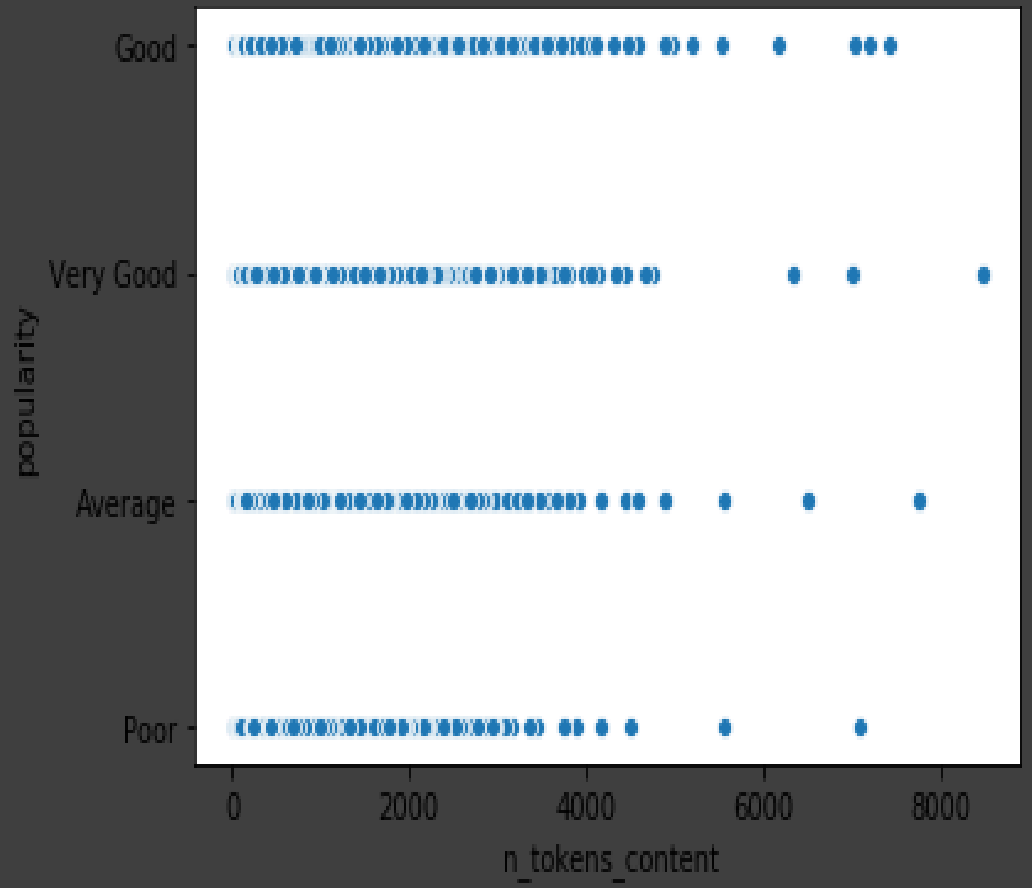
TABLE 1: POPULARITY CLASS CLASSIFICATION

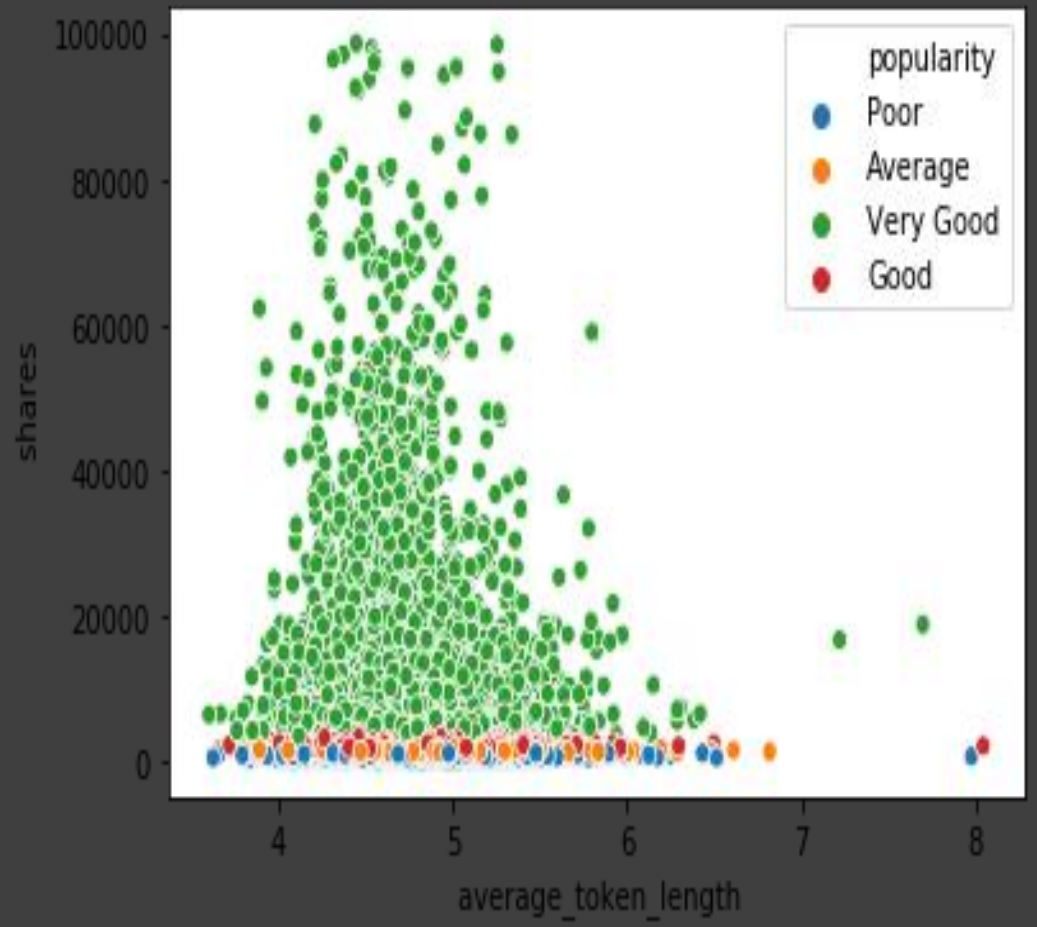
Class distribution for Insight analysis



Quantitative Analysis of Hypothesis

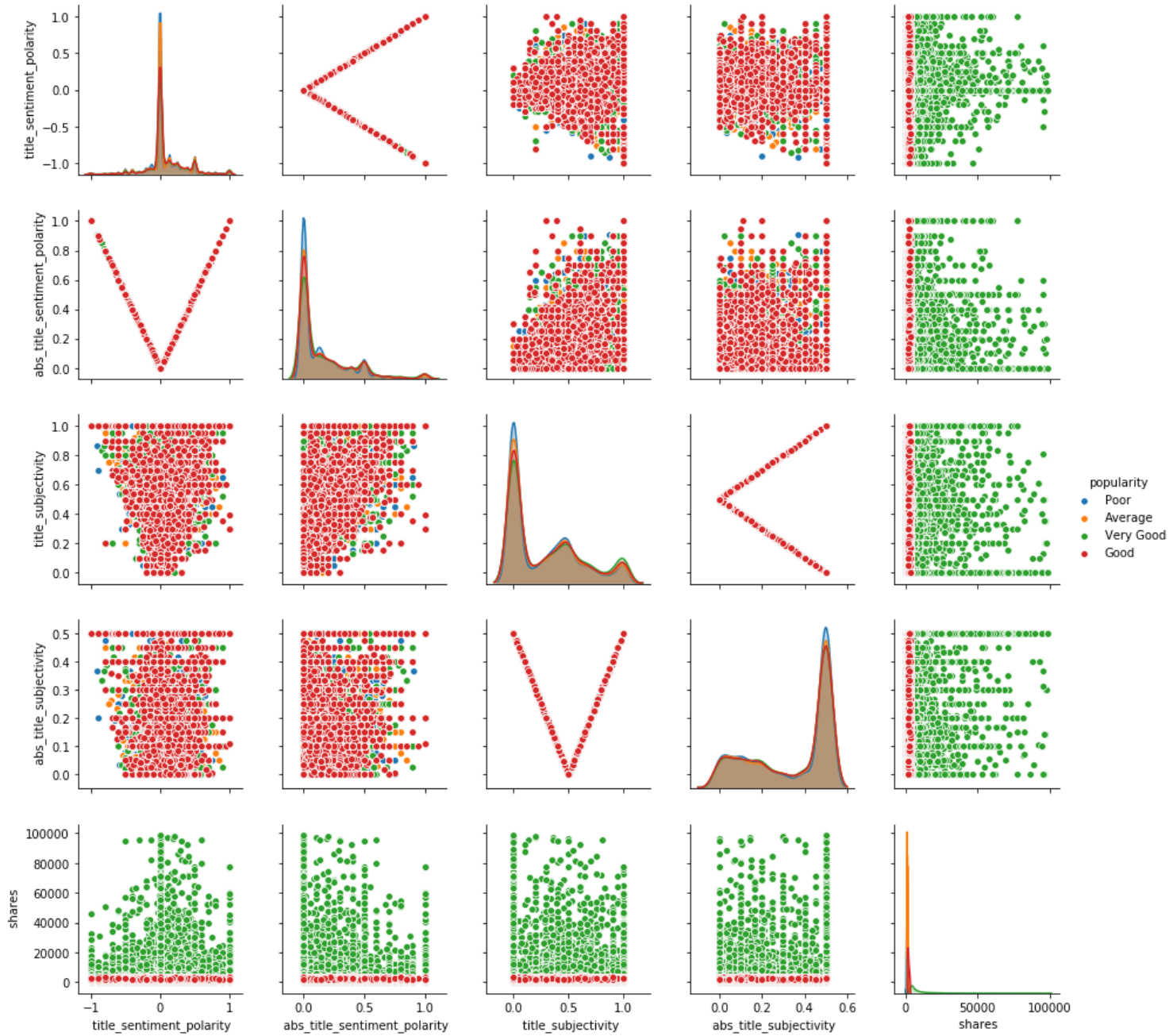
- Subjective analysis is a good way to kickstart a project, but it might not be enough.
- We embark on carrying both Univariate and Bivariate analysis on variables in the dataset to confirm or debunk our hypothesis.



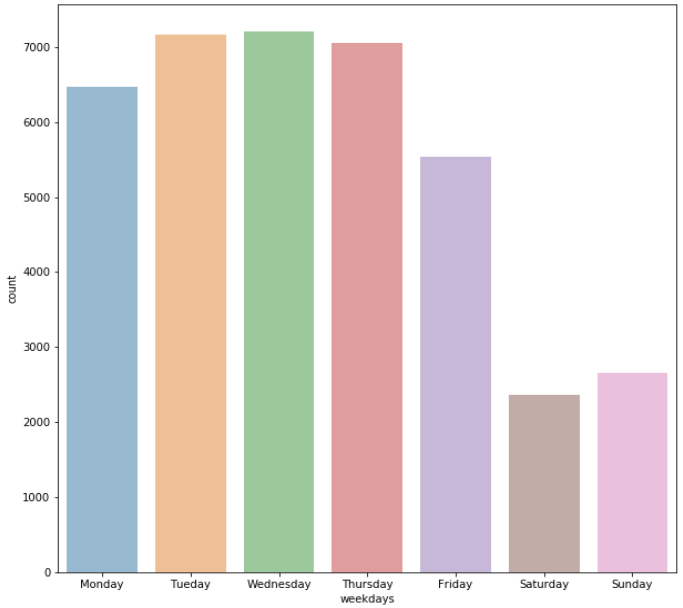


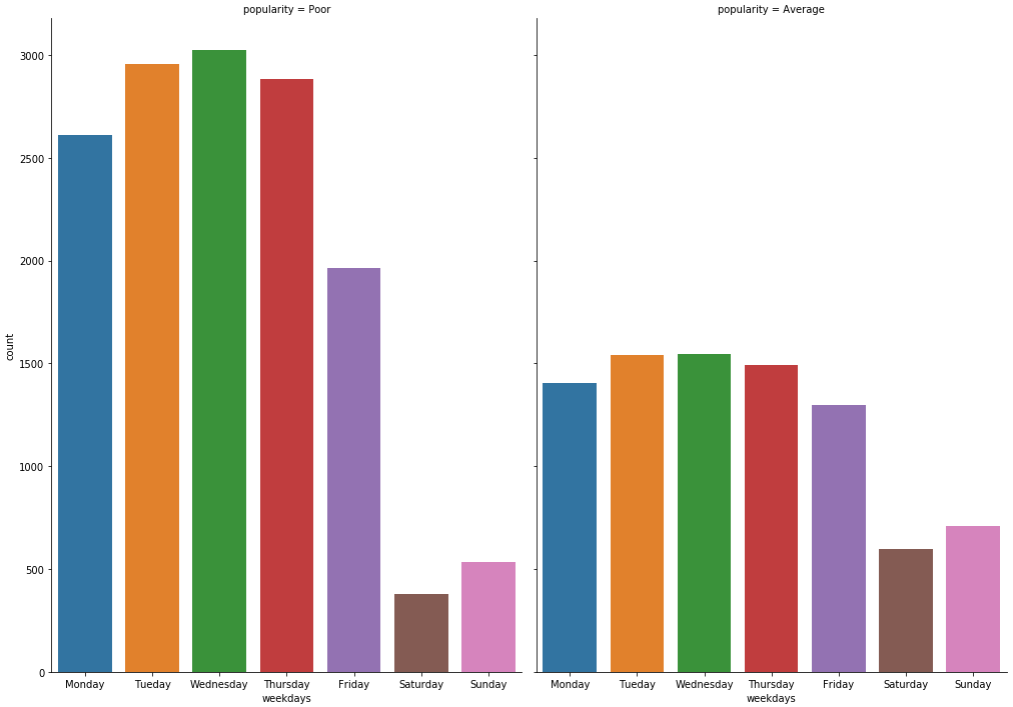
Data Visualization - tokens

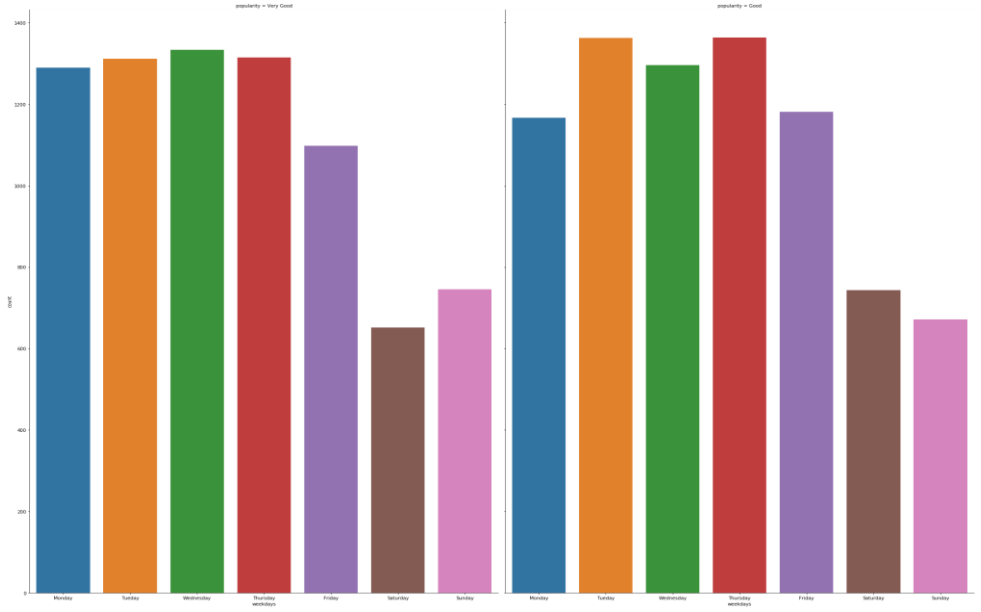
Data Visualization



Weekdays







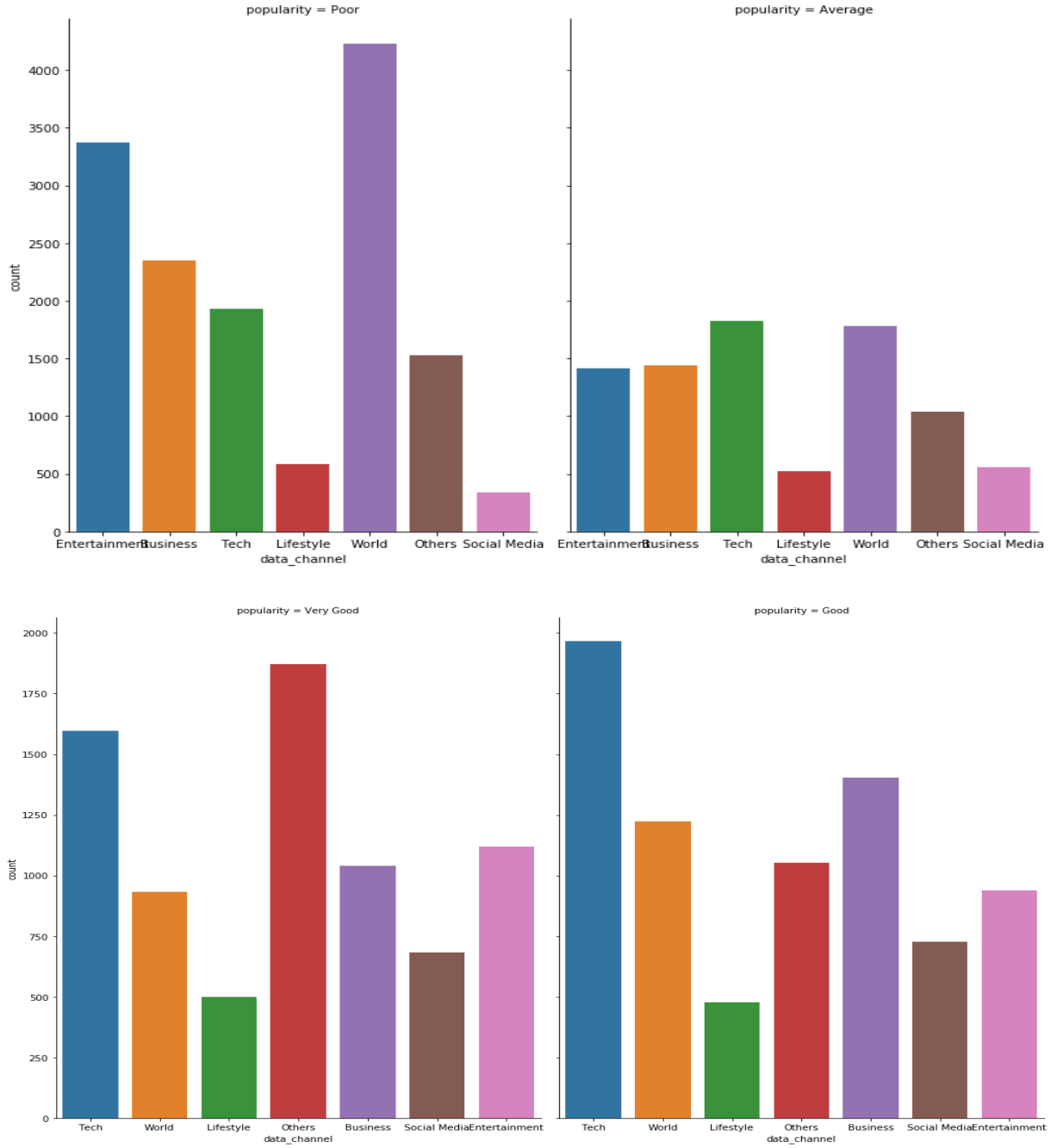
Most of the articles in Mashable were published on weekdays as compared as compared to weekends.

It seems the best popular articles are usually posted on Mondays and Wednesday (and a bit of Tuesdays) Sundays and Saturdays (Weekends generally) are the worsts days to publish an articles. Your chances are low

Channels

Most articles are published in World, Technology and Entertainment data channel.

Best articles with highest popularity belongs to the others, business and entertainment data channel.



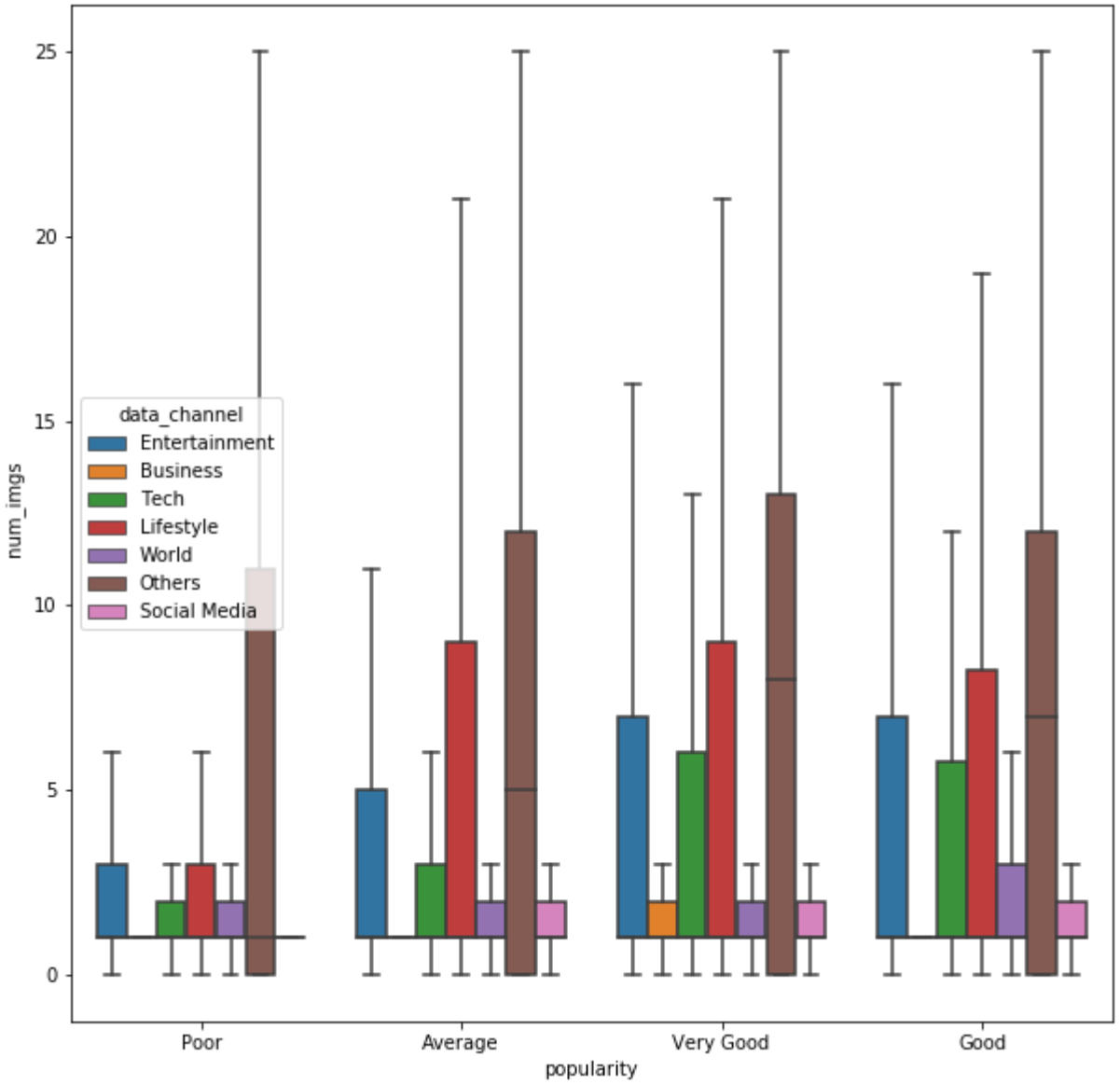
Bi-variate analysis

Num_images vs Data Channel vs Popularity

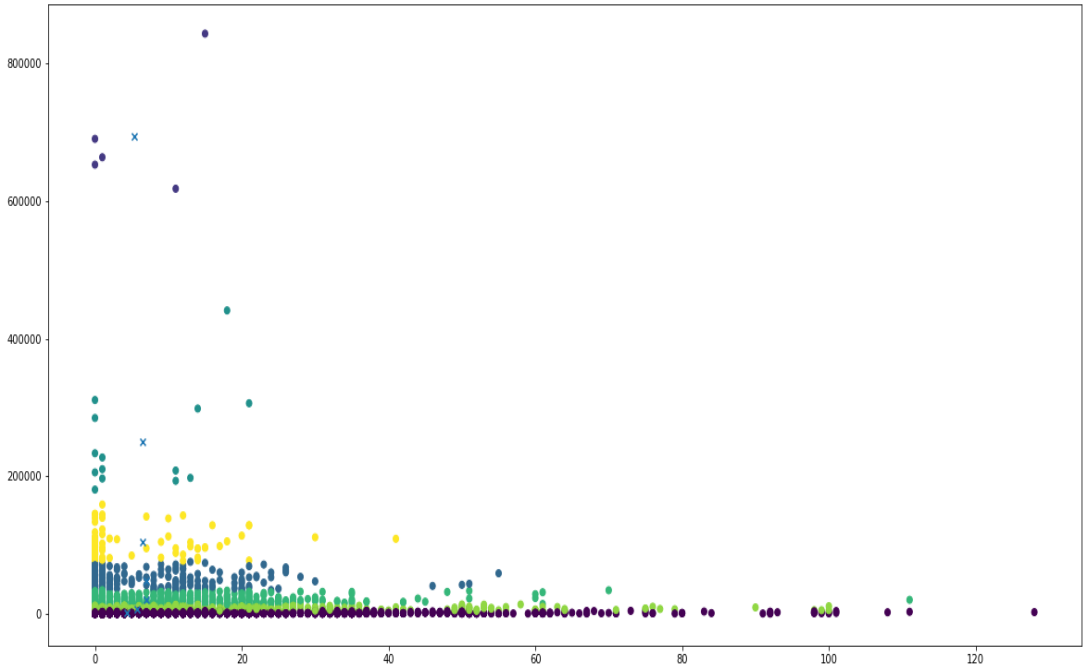
Popular articles tend to have higher visuals in them, but it is not always the case.

Business channels generally don't get influenced by the num_images in them. They generally have low images irrespective of its popularity.

This is a peculiar pattern and Entertainment channels generally tend to have high visuals as their popularity increases.

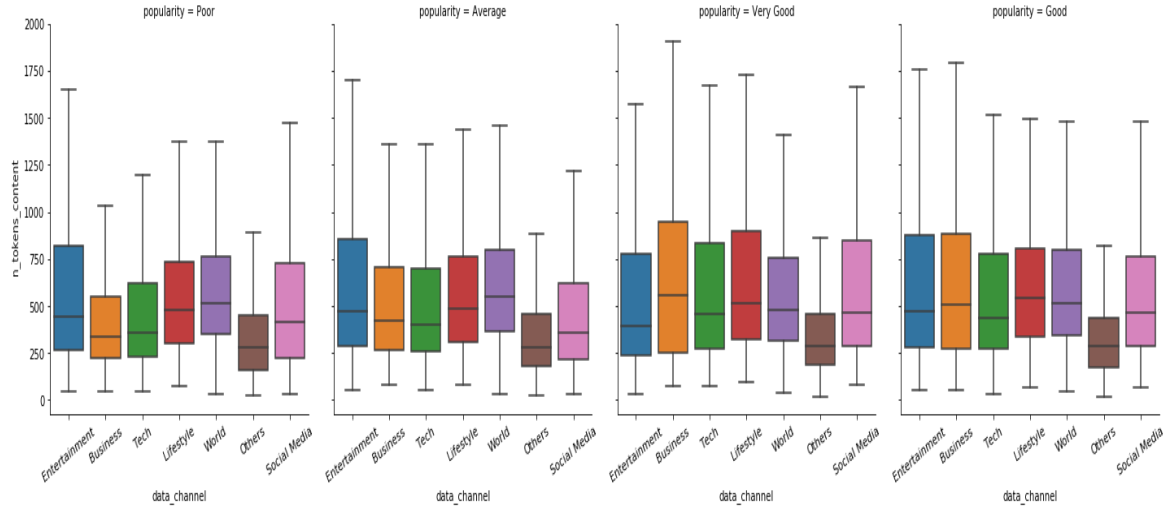


- It can also be seen through Kmeans clustering that Business data channel has popular article(High shares) with less number of images



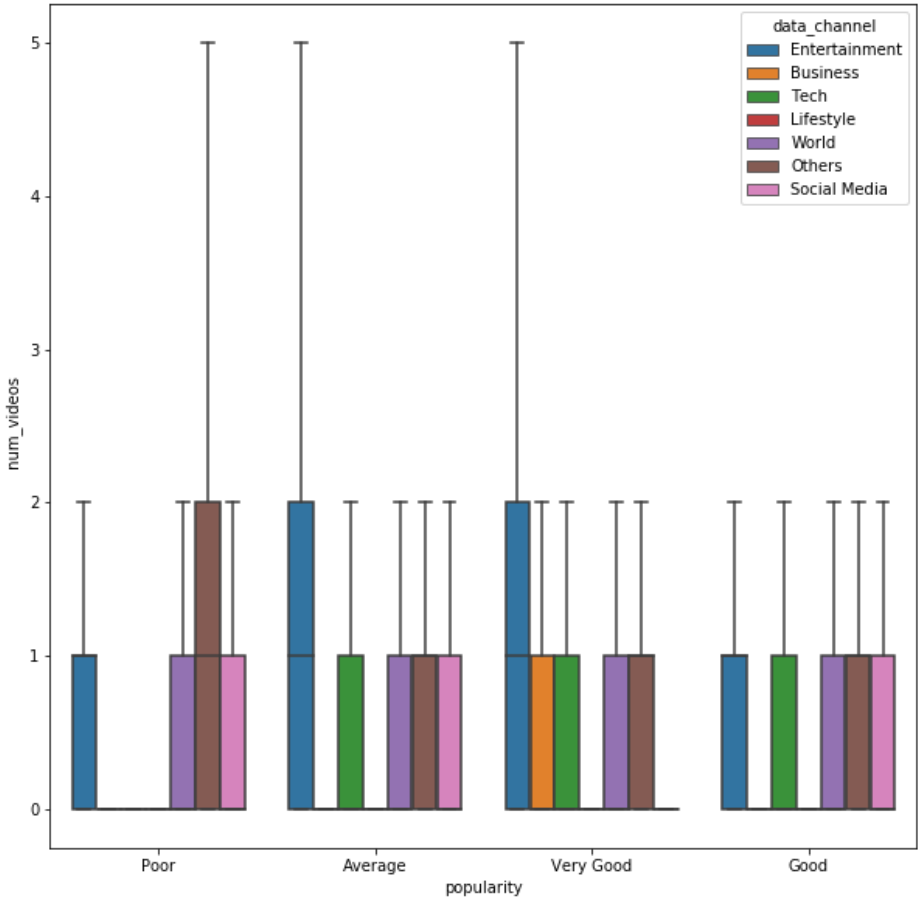
N_tokens_content vs data channel

- From our previous observation we had said that lesser the content popular the article.
- Plot here shows that good data channel like Entertainment and Business have higher content compared other Data channel.



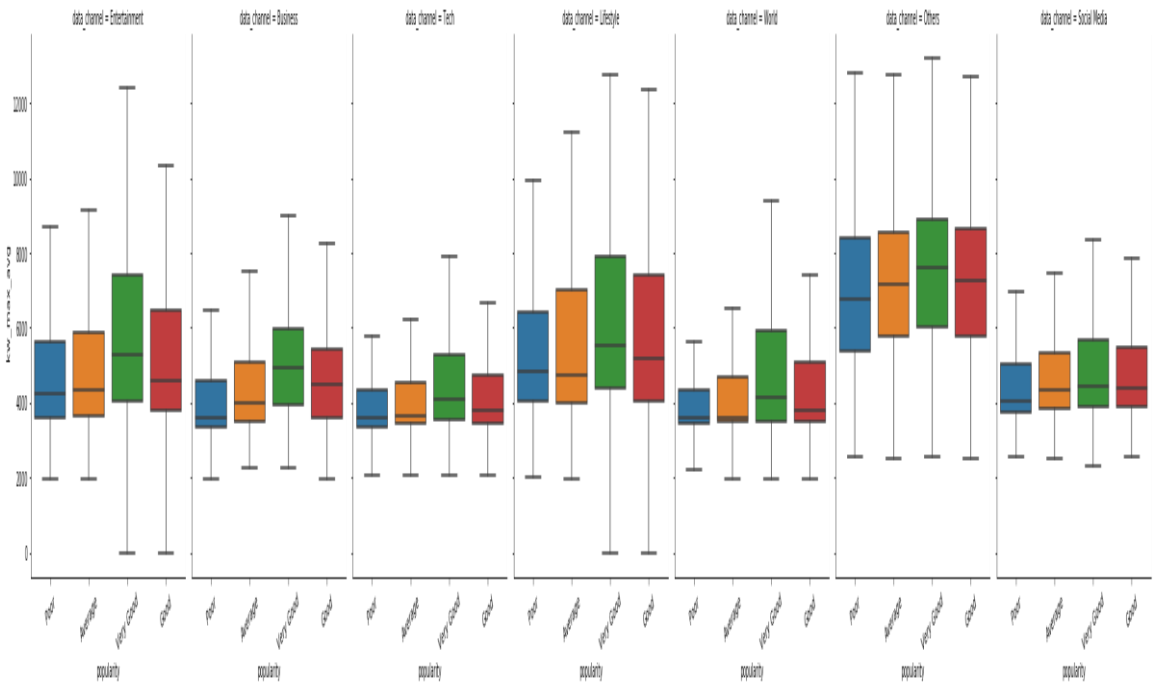
Num_videos vs data_channel vs popularity

- From univariant analysis we got to know that having more videos makes an articles less popular.
- But from this plot it can be seen that Entertainment articles tend to have more number of videos in popular categories and Business channel have lesser number of videos.



Impact of kw_max_avg on data_channel

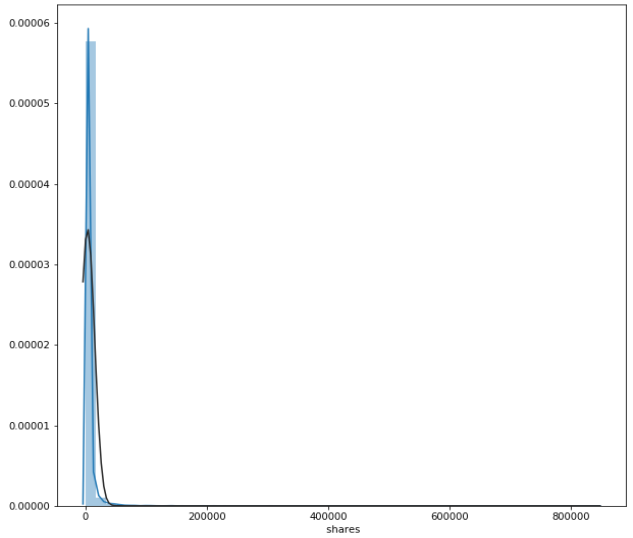
- Popular data channel like Entertainment, Business and Technology tend to have lower average kw_max_avg value compared to other data channels, which is opposite to what is expected from popular data channel

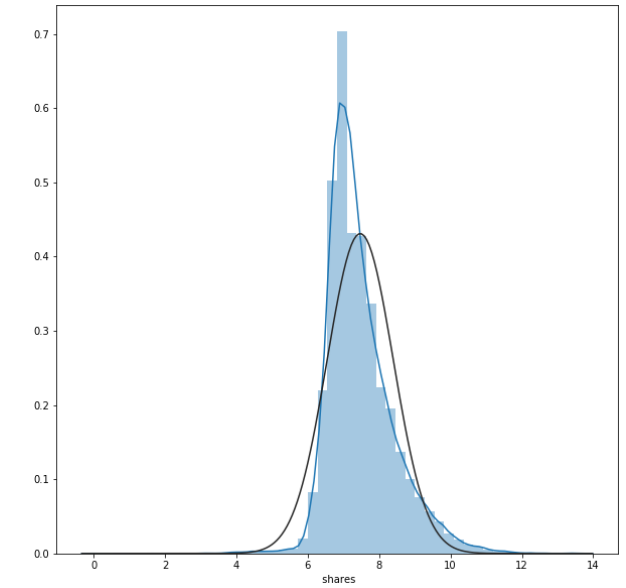


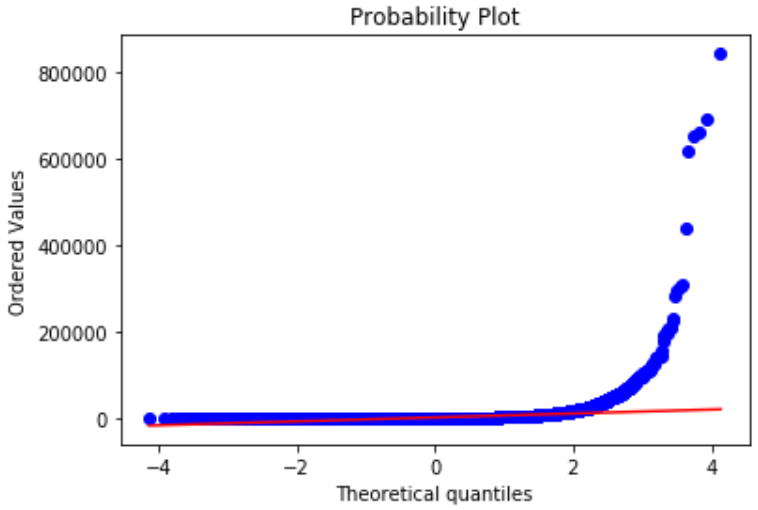
Hypothesis Update

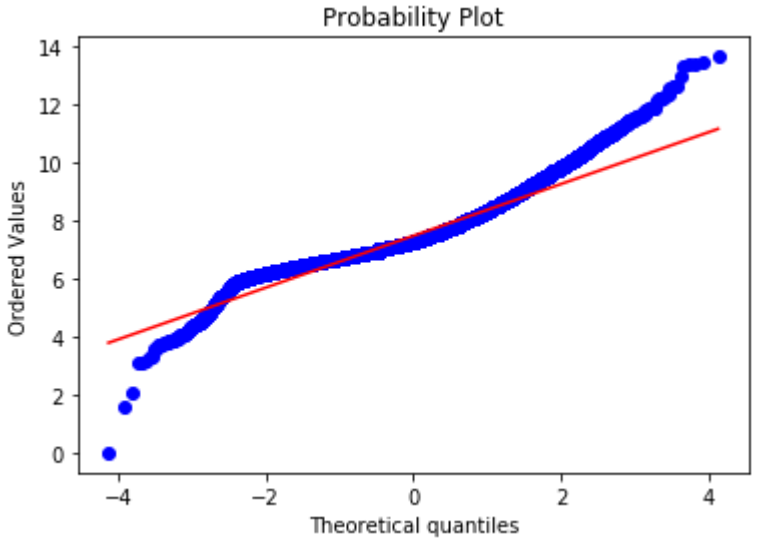
Low	Medium	High
n_non_stop_words, n_non_stop_unique_tokens kw_min_min kw_min_max kw_min_avg kw_max_min kw_max_max kw_max_avg kw_avg_min LDA_00 LDA_01 LDA_02 LDA_03 LDA_04 rate_positive_words rate_negative_words min_positive_polarity max_positive_polarity min_negative_polarity max_negative_polarity abs_title_subjectivity abs_title_sentiment_polarity	n_tokens_title n_unique_tokens num_hrefs num_self_hrefs num_imgs num_videos average_token_ length kw_avg_max kw_avg_avg self_reference_min_shares self_reference_max_shares self_reference_avg_shares global_subjectivity global_sentiment_polarity global_rate_positive_words global_rate_negative_words avg_positive_polarity avg_negative_polarity	n_tokens_content num_keywords data_channel Weekdays title_sentiment_polarity

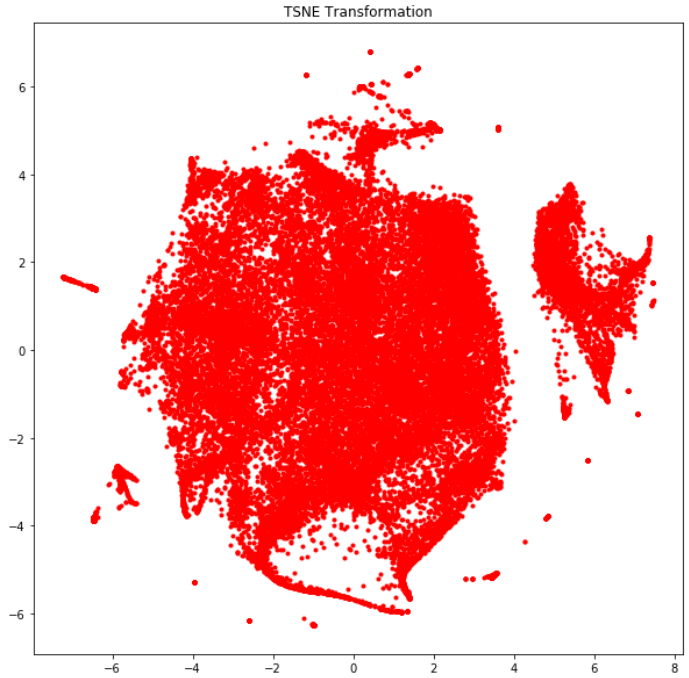
Normal Distribution

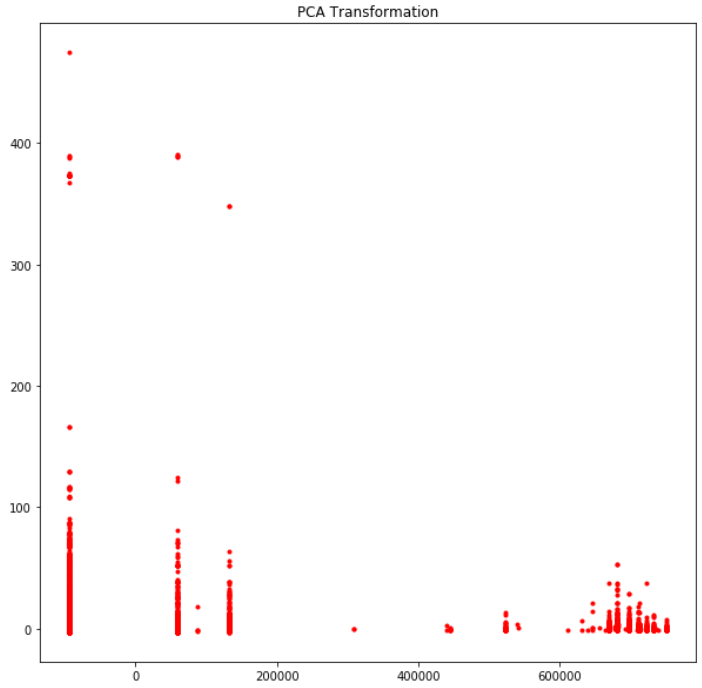






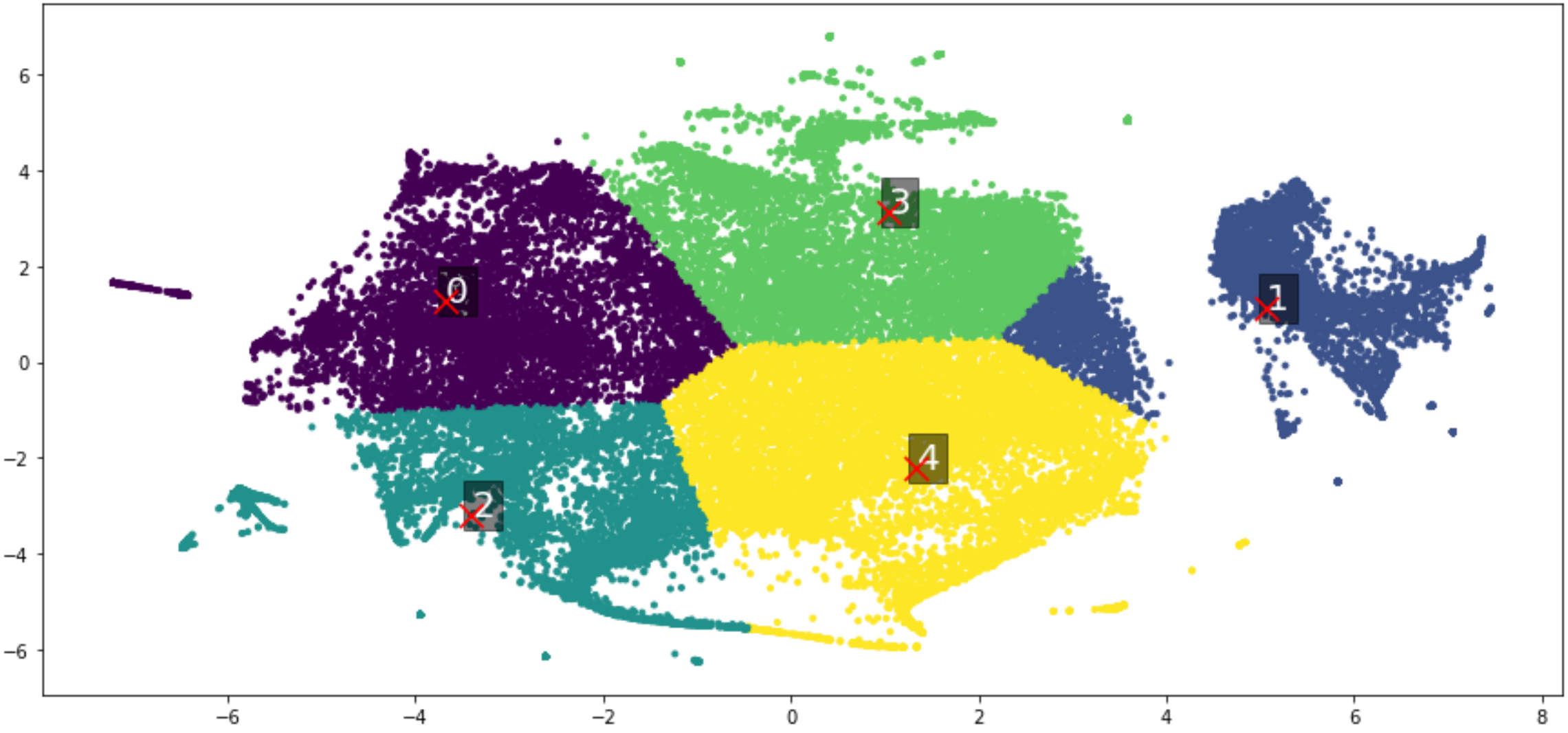






Cluster Extraction

- This was done to find special pattern from data and group the similar articles into clusters that have similar traits.
- Clustering was done using KMeans
- Two dimensionality reduction approach was considered: Principal Component Analysis
- 1. PCA
- 2. T-distributed stochastic neighbor embedding (T- TSNE)



Cluster formation by Kmeans Algorithm

FEATURE SELECTION AND EXTRACTION

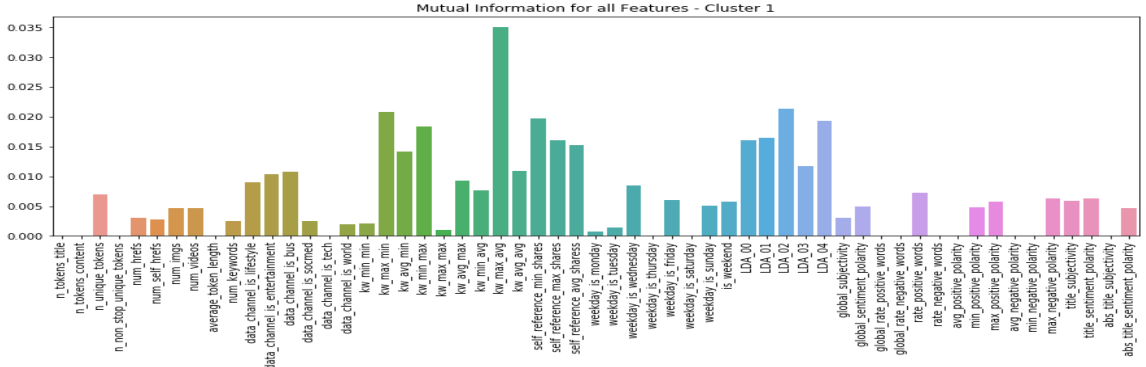
We have used four feature selection techniques

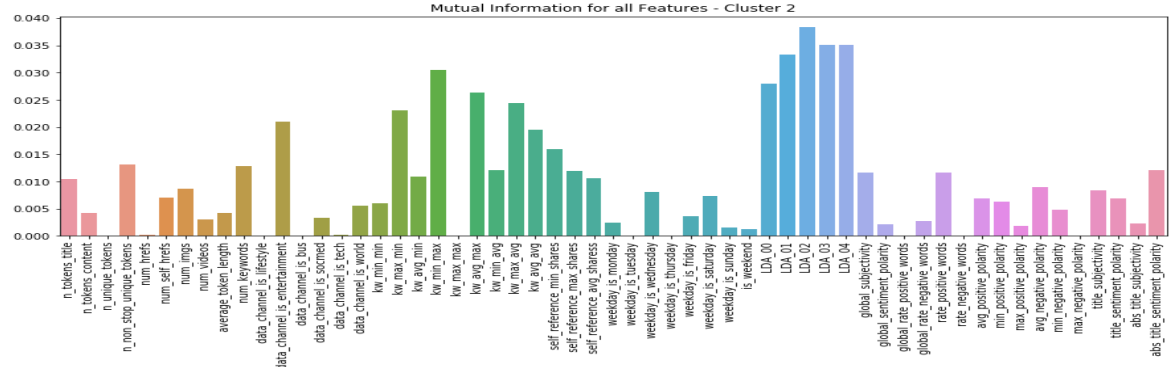
- Mutual Information
- F-Score
- Recursive Feature Selection
- PCA

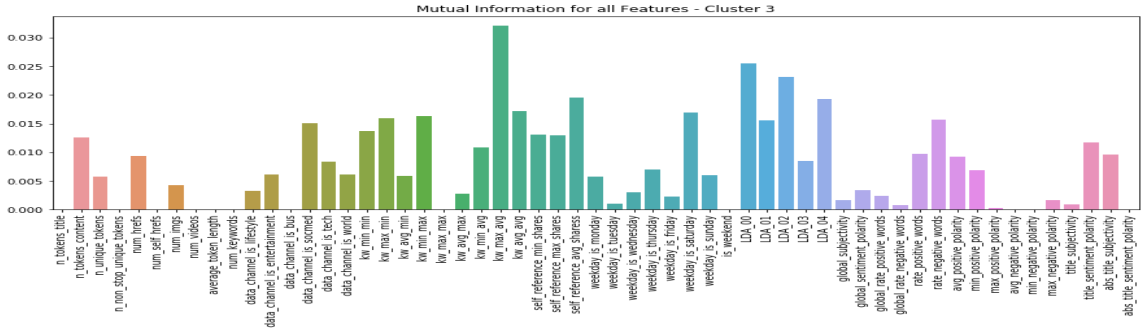
We have used four feature space selection

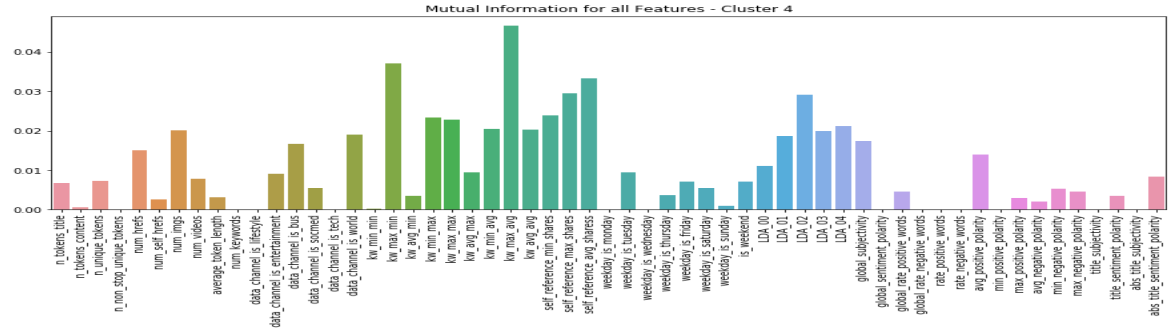
- Top 5
- Top 10
- Top 20
- Top 30

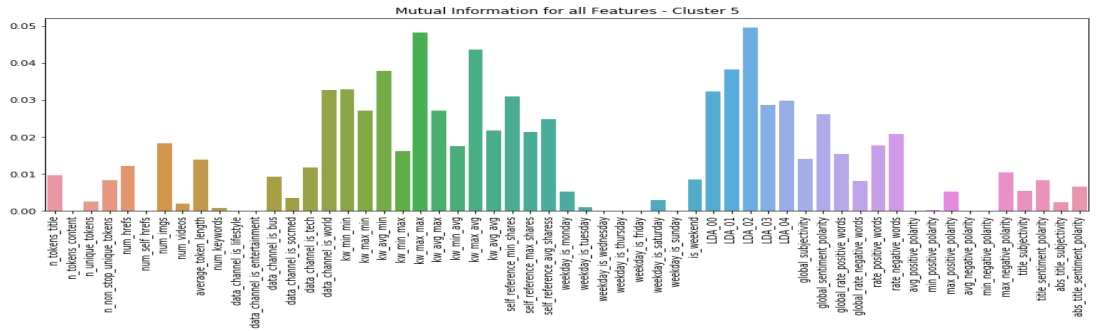
Barplot- Features vs MI Score



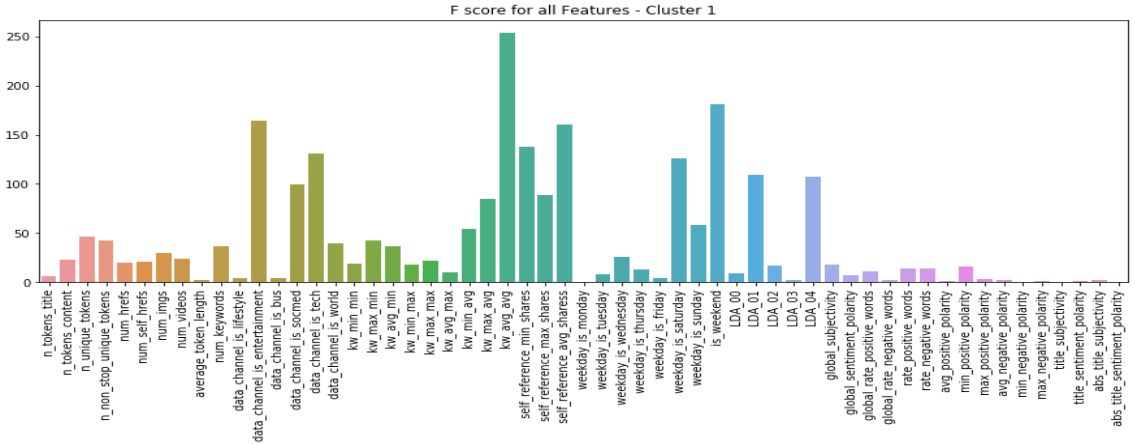


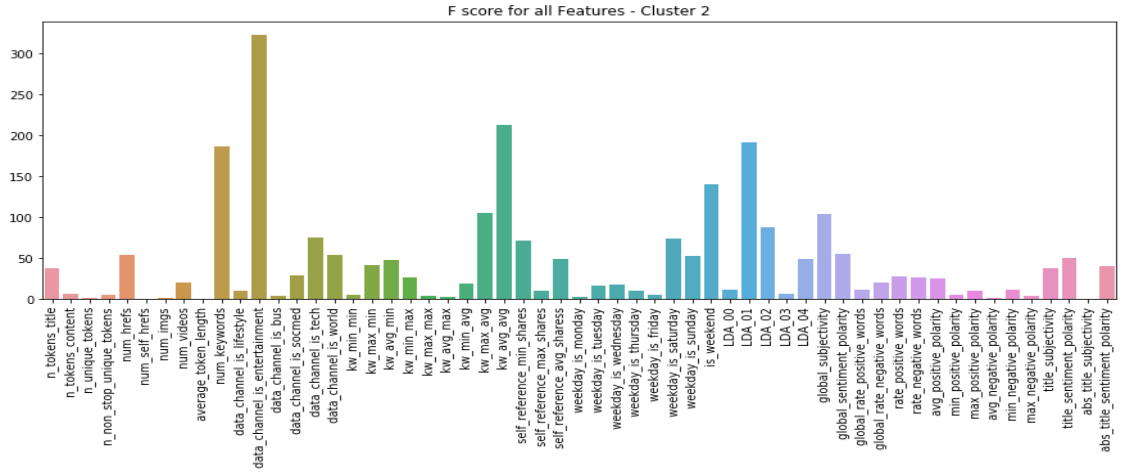


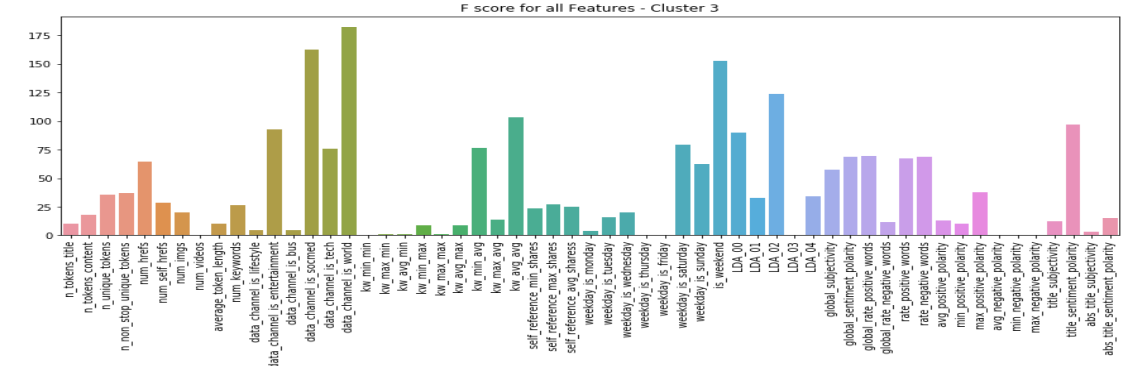


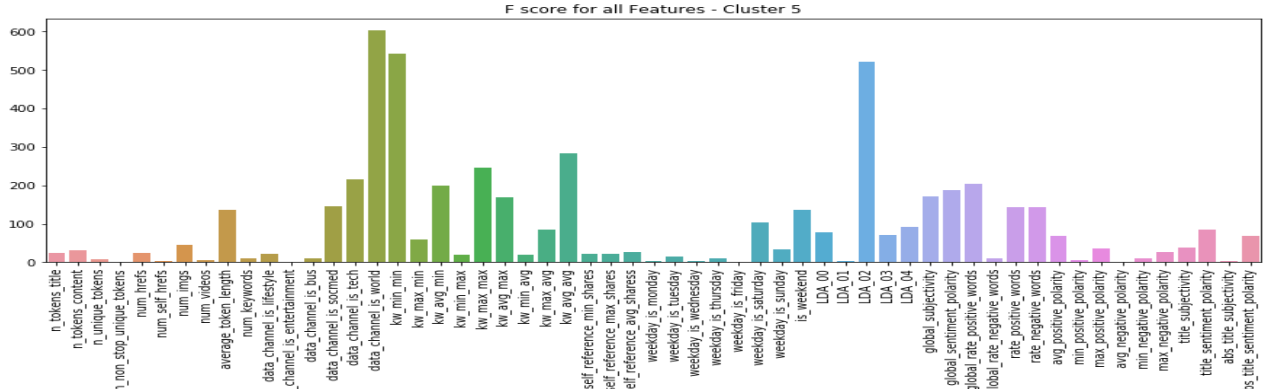


Bar plot- Features vs F-Score





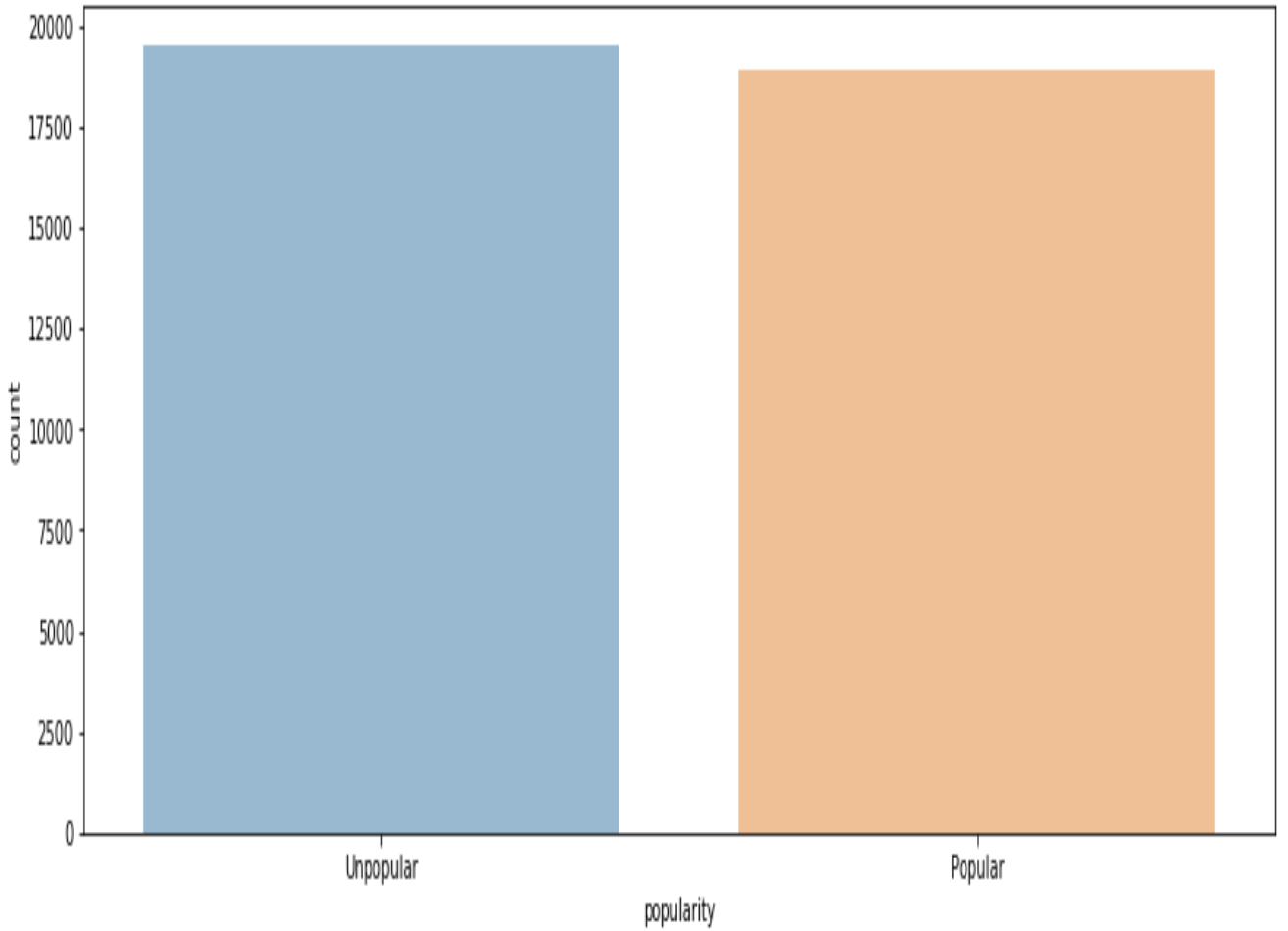




Machine Learning Models

Three Machine Learning Models were with binary target is considered:

- KNN
- Random Forest
- SVM



	popularity	No of articles
0	Popular	18911
1	Unpopular	19551

Result and Experiments

- Table below shows cluster performance with top features from feature selection techniques

Cluster	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
Cluster 1	63.3	63.67	64.51	**64.99**
Cluster 2	67.08	67.39	**70.01**	68.82
Cluster 3	**68.82**	**68.82**	**68.82**	**68.82**
Cluster 4	**68.82**	**68.62**	**68.82**	**68.82**
Cluster 5	73.65	74.02	**74.55**	74.41

Cluster 1 Evaluation

	Method	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
PCA	KNN	58.44	60.82	**62.88**	61.24
	Random Forest	56.81	60.13	62.19	**62.82**
MI	KNN	61.03	62.49	**62.56**	61.82
	Random Forest	61.19	61.61	63.41	**64.30**
F-Score	KNN	63.30	63.67	64.04	**64.78**
	Random Forest	58.98	62.14	63.83	**64.78**
RFE	KNN	62.24	62.35	**62.51**	62.24
	Random Forest	61.08	63.51	64.51	**64.46**
	SVM	62.08	64.36	63.83	**64.99**

Cluster 2 Evaluation

	Method	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
PCA	KNN	59.85	62.09	62.84	**63.27**
	Random Forest	58.56	63.50	64.63	**66.04**
MI	KNN	59.28	63.15	**64.46**	63.59
	Random Forest	60.97	63.71	68.14	**68.82**
F-Score	KNN	66.27	**67.14**	66.70	64.83
	Random Forest	61.47	64.96	70.01	67.95
RFE	KNN	66.58	**66.95**	66.45	63.02
	Random Forest	64.15	67.01	67.45	**68.32**
	SVM	67.08	67.39	**67.95**	67.08

Cluster 3 Evaluation

	Method	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
PCA	KNN	57.49	57.43	60.11	60.70
	Random Forest	57.77	60.65	61.92	**64.10**
MI	KNN	60.90	60.24	60.96	**61.62**
	Random Forest	**68.82**	**68.82**	**68.82**	**68.82**
F-Score	KNN	61.29	60.37	**62.14**	61.68
	Random Forest	55.46	60.11	62.67	**63.06**
RFE	KNN	62.34	**62.80**	60.37	60.83
	Random Forest	59.85	63.52	64.57	**64.63**
	SVM	61.29	61.62	**63.26**	61.55

Cluster 4 Evaluation

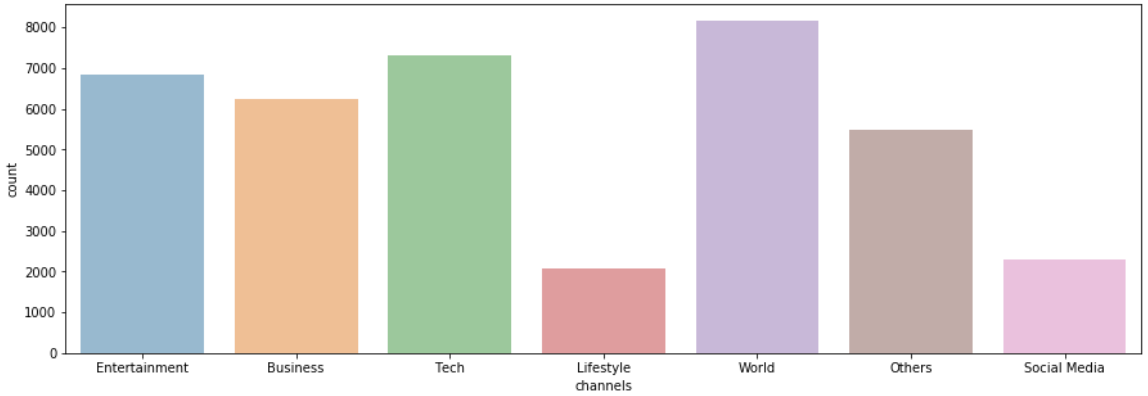
	Method	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
PCA	KNN	64.39	65.46	65.21	66.03
	Random Forest	57.23	58.27	58.99	61.95
MI	KNN	65.87	63.98	64.8	64.55
	Random Forest	68.82	68.82	68.82	68.82
F-Score	KNN	65.87	68.25	65.70	64.80
	Random Forest	61.10	66.2	68.66	68.42
RFE	KNN	67.59	67.43	66.61	66.36
	Random Forest	65.04	67.43	68.75	68
	SVM	67.26	67.76	67.18	66.36

Cluster 4 Evaluation

	Method	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
PCA	KNN	64.39	65.46	65.21	66.03
	Random Forest	57.23	58.27	58.99	61.95
MI	KNN	65.87	63.98	64.8	64.55
	Random Forest	68.82	68.82	68.82	68.82
F-Score	KNN	65.87	68.25	65.70	64.80
	Random Forest	61.10	66.2	68.66	68.42
RFE	KNN	67.59	67.43	66.61	66.36
	Random Forest	65.04	67.43	68.75	68
	SVM	67.26	67.76	67.18	66.36

Cluster 5 Performance

	Method	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
PCA	KNN	73.10	73.93	74.34	**74.41**
	Random Forest	67.17	69.32	71.34	71.25
MI	**KNN**	73.65	73.45	**74.55**	74.41
	Random Forest	68.82	68.82	68.82	68.82
F-Score	KNN	73.17	74.02	74.02	**74.36**
	Random Forest	67.88	72.28	72.76	**73.38**
RFE	KNN	73.17	72.83	74.41	**74.48**
	Random Forest	70.15	70.90	73.03	**74.48**
	SVM	73.52	73.52	**73.86**	73.65



Data Channel	No.of Articles
Business	6235
Entertainment	6855
Lifestyle	2077
Others	5491
Social Media	2311
Technology	7325
World	8168

DATA CHANNEL

- Goal: Predict data channel of an article
- One of the contributing feature

Results- Data Channel

- All features (whole dataset) was introduced, Random Forest gave accuracy of 82.19%
- Table below shows best result of each cluster along with feature space

Cluster	Top 5(%)	Top 10(%)	Top 20(%)	Top 30(%)
Cluster 1	77.54	**80.22**	79.92	80
Cluster 2	**89.74**	88.78	88.92	88.85
Cluster 3	76.26	**79.51**	78.63	75.57
Cluster 4	80.20	82.96	**83.40**	82.81
Cluster 5	73.71	75.71	75.99	**77.06**

- Cluster 1 Result

	Method	Top 5 (%)	Top 10 (%)	Top 20 (%)	Top 30 (%)
f-score	KNN	74.49	**75.01**	73.01	70.11
	Random Forest	77.54	**80.22**	79.92	80

- Cluster 2 Result

	Method	Top 5 (%)	Top 10 (%)	Top 20 (%)	Top 30 (%)
f-score	**KNN**	**89.74**	88.48	84.87	83.09
	Random Forest	**88.92**	88.78	**88.92**	88.85

- Cluster 3 Result

	Method	Top 5 (%)	Top 10 (%)	Top 20 (%)	Top 30 (%)
f-score	KNN	72.22	**73.05**	68.12	65.23
	Random Forest	76.26	**79.51**	78.63	75.57

- Cluster 4 Result

	Method	Top 5 (%)	Top 10 (%)	Top 20 (%)	Top 30 (%)
f-score	KNN	72.22	**73.05**	68.12	65.23
	Random Forest	76.26	**79.51**	78.63	75.57

- Cluster 5 Result

	Method	Top 5 (%)	Top 10 (%)	Top 20 (%)	Top 30 (%)
f-score	KNN	76.26	**79.51**	78.63	75.57
	Random Forest	80.20	82.96	**83.40**	82.81

Conclusion

Quantitative analysis was used to confirm our initial hypothesis

Two popularity classes were considered for popularity prediction and unsupervised learning was used to transform to two-dimensional data

Machine learning model was built to be able to predict the popularity class and data channel

The best machine learning model obtained was Random Forest, which was able to attain an accuracy of 75% for popularity prediction and 89% for data channel prediction

Recommendations

- The number of words in the article should be less than 1500 words.
- Article title should have the right length (6 – 17)
- Articles should have a good amount of images. Between 1 – 40 images are great. Having higher number of keywords and unique words helps in achieving better popularity
- Increase the amount of subjectivity in the title and content.
- Publishing articles focusing on treading topics.

Thank You