ONLINE NEWS POPULARITY

Nadir Khan

Lucas Di Monte

York University

Big Data Analytics Capstone project

1. Business problem

1.1. Business Understanding

Source: http://www.danpontefract.com/employee-access-to-social-media-in-the-workplace-decreases/

The balance of the world economy has been shifting; thereby allowing billions of people in Asia, South America and Africa to move out of poverty and into the middle-income classes?

This emergence of the middle-class has resulted in higher accessibility to the internet. This access has allowed more people to connect globally using social media. Currently, 2.62 Billion people are on social media and it is expected to increase to 3.02 billion (a third of the world's population) by 2021.

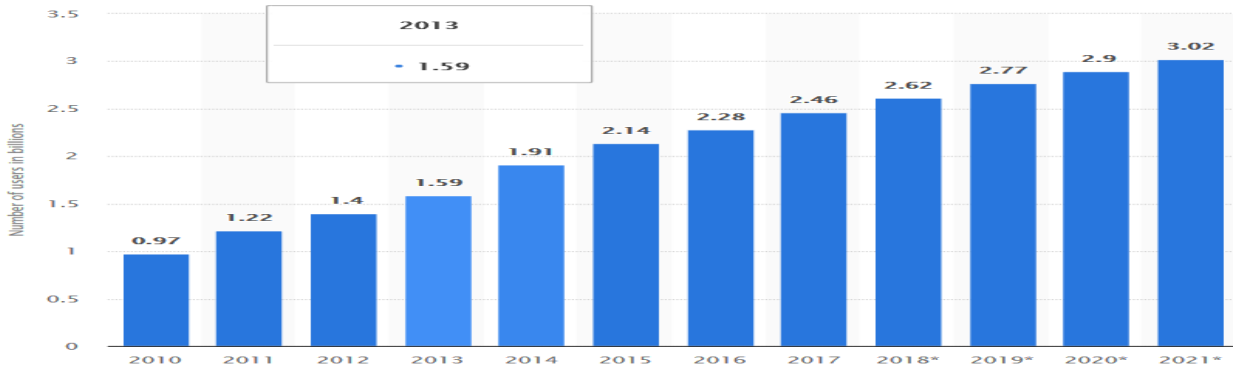


Figure 1 shows growing number of online users over the years

Source: Worldwide; eMarketer;

Social media has increased interest in online news, as it allows information to be spread rapidly globally. Thus, predicting the popularity of online news is becoming a recent research trend. Popularity is often measured by considering the number of shares in social networks, which is valuable for advertisers and accused of influencing political campaigns i.e. Cambridge Analytics.

The first survey and research report to measure "Social Media in the Workplace Global Study" was prepared by a Proskauer, a global law firm providing a wide variety of legal services to clients worldwide.

The 2014 version of the survey provided the following interesting facts regarding social media:

- 90% of companies now use social media for business purposes
- Social media policies are now found in 80% of organizations, up from 60%
- Only 17% of organizations have provisions that protect them against the misuse of social media by ex-employees
- 36% of employers actively block access to such sites, compared to 29% in 2012
- 43% of businesses permit their employees to access social media sites

Source: http://www.danpontefract.com/employee-access-to-social-media-in-the-workplace-decreases/

Based on the above facts, our team will predict the popularity of an online news article prior to being published by using a dataset prepared by Mashable Inc. We considered an article as "popular" if the number of shares is higher than a fixed decision threshold, else it is considered "unpopular". These results could help users to predict the popularity of their articles to determine the amount to charge for advertisements.

2. Analytical problem

2.1. Understanding of Data

2.1.1. Sources

The "Online News Popularity" dataset was compiled by a Mashable Inc., a digital media website founded in 2005, with over 9.5 million Twitter followers and over 6.5 million fans on Facebook. This site publishes hundreds of articles daily on topics such as lifestyle, social media, business, news, entertainment, sports, technology. We imported these data from these articles from the CSV file found on the UCI Machine Learning Repository. (https://archive.ics.uci.edu/ml/datasets/online+news+popularity#)

2.1.2. Data Dictionary

The dataset has 61 attributes describing each aspect of the articles and 39,797 number of observations. We used the following categorization table provided by Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez.

Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)
Article text polarity score and its absolute difference to 0.5	ratio (2)

Target	Type (#)
Number of article Mashable shares	number (1)

Figure 2 shows the data set structure

Source: https://pdfs.semanticscholar.org/ad7f/3da7a5d6a1e18cc5a176f18f52687b912fea.pdf

3. Exploratory data analysis

Shares by Average Negative Polarity:

The number of shares is highest when the Average Negative Polarity is around – 0.25.

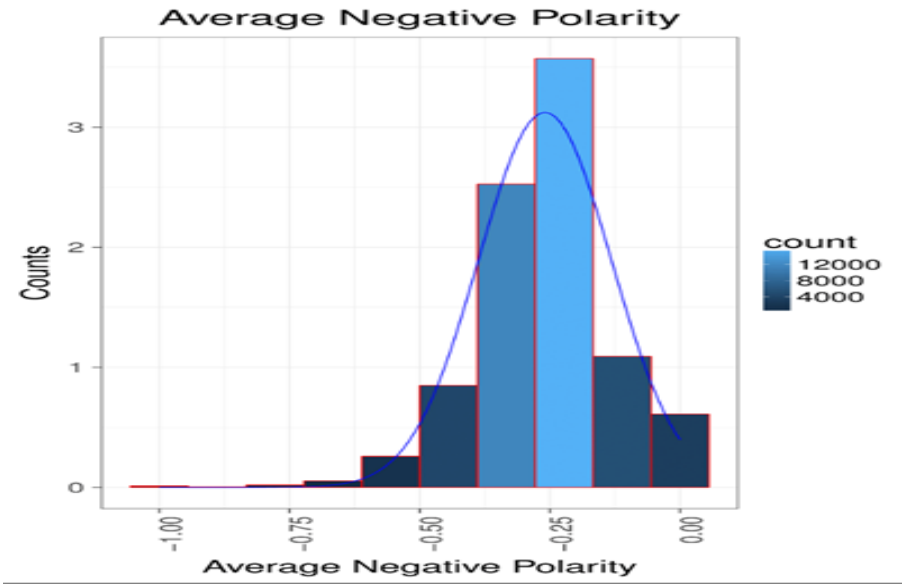
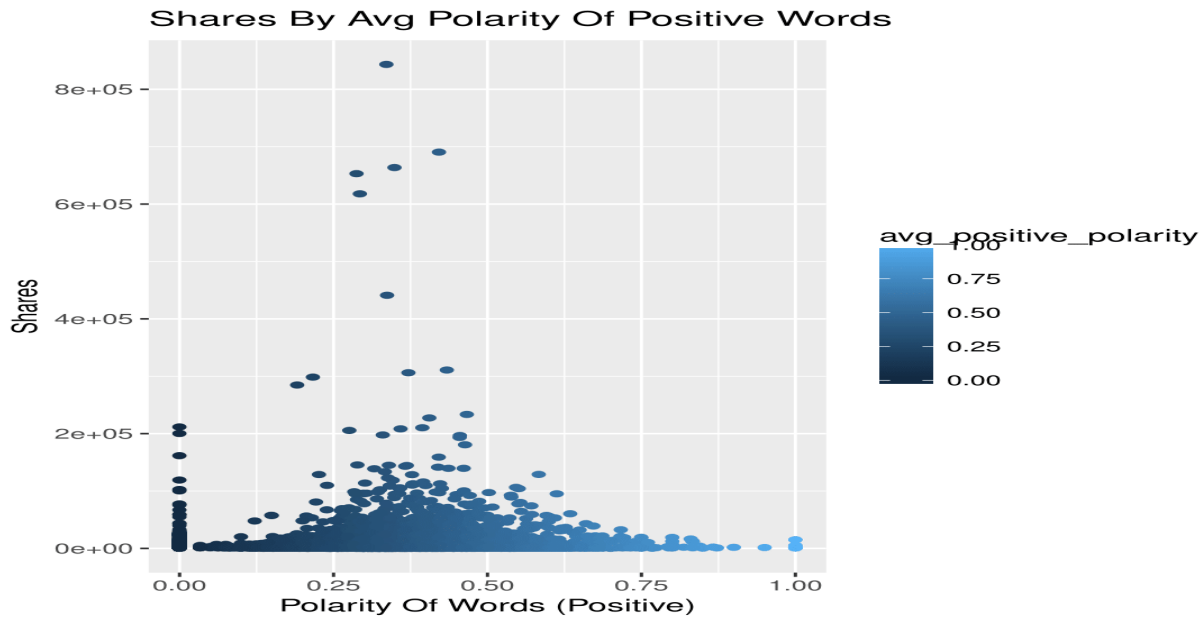


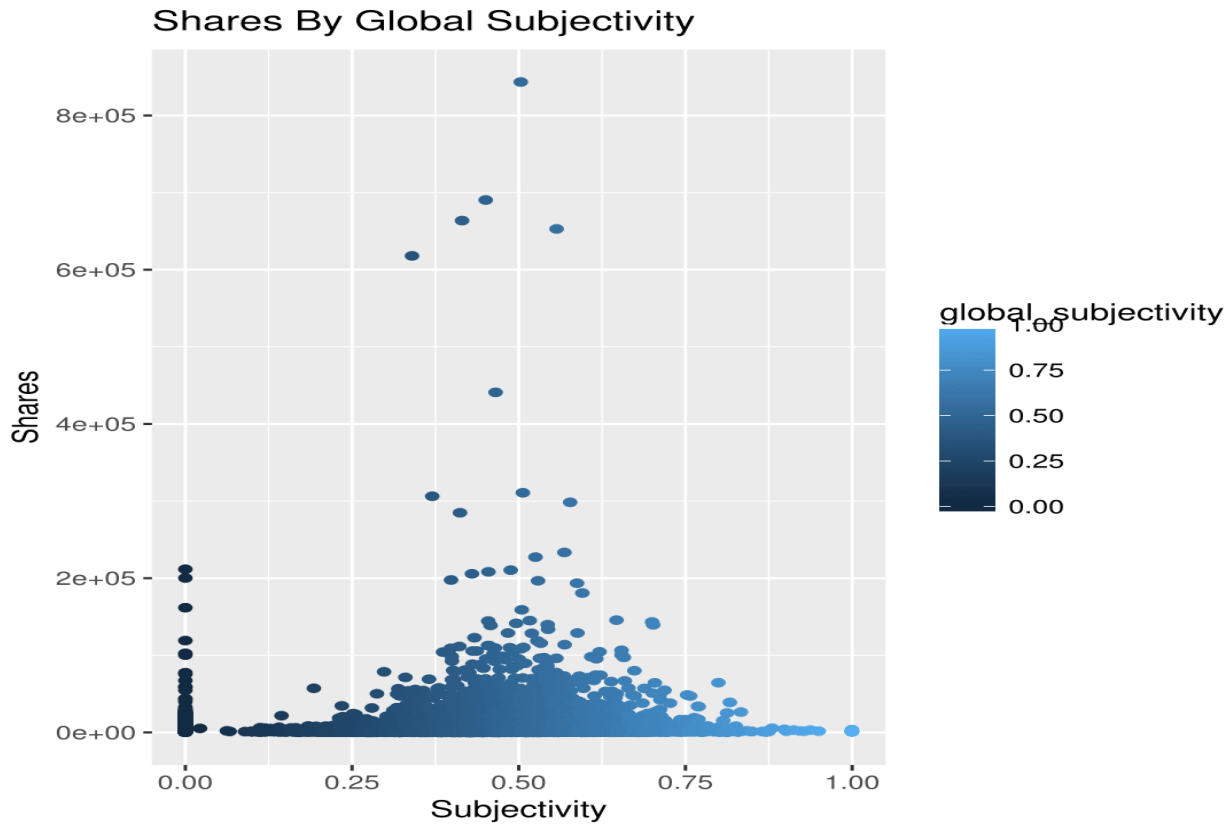
Figure 3 shows the counts of average negative polarity

Shares by Average Polarity of Positive Words: As we go from 0.25 to 0.50, we determine that this is when the shares are the most.

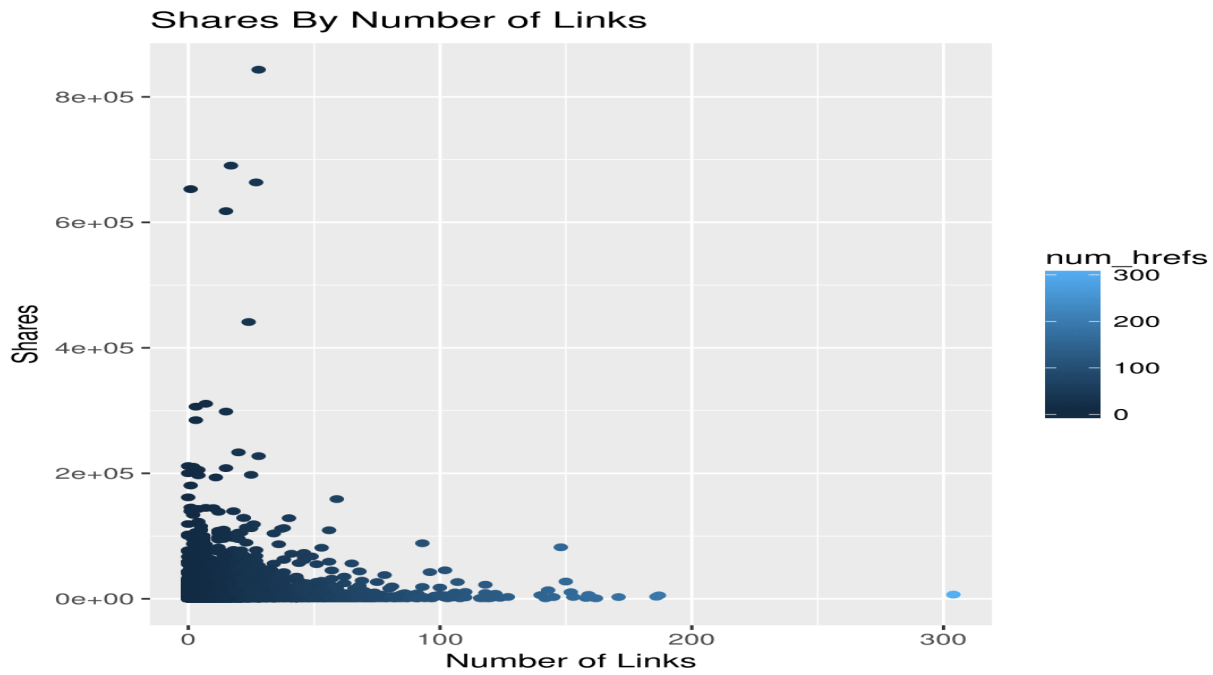


Subjectivity:

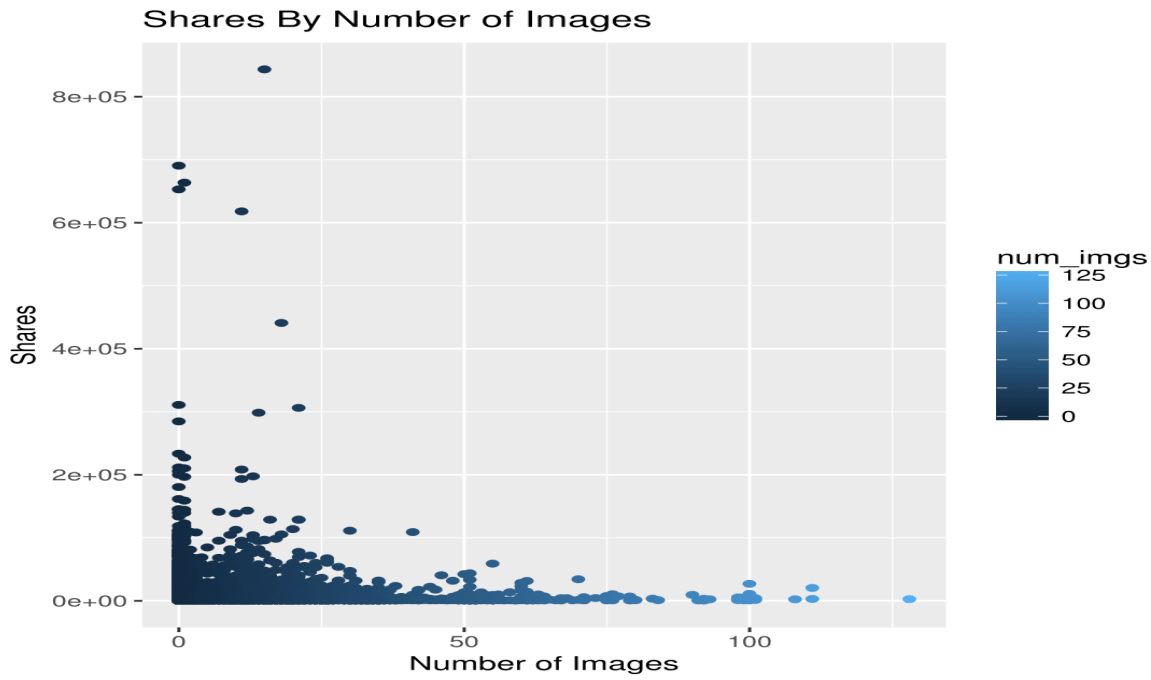
The articles with the most near average subjectivity are in the .25 to .75 range.



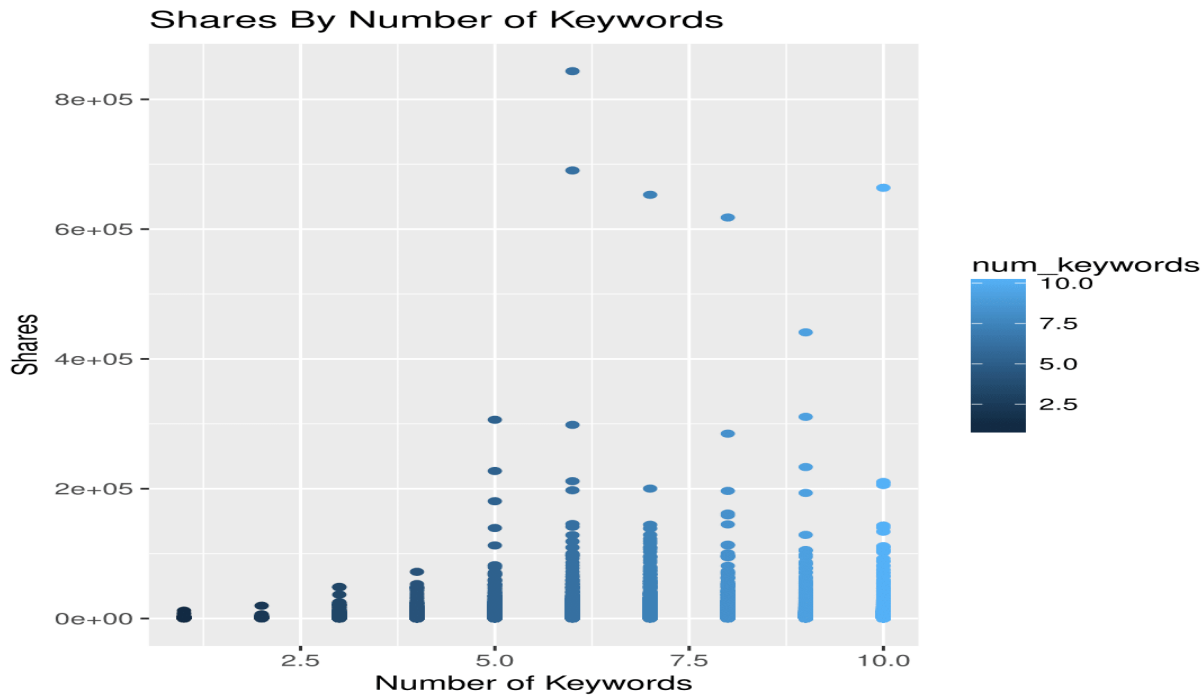
Share by number of links:
The higher the number of links will the decrease the number of shares.



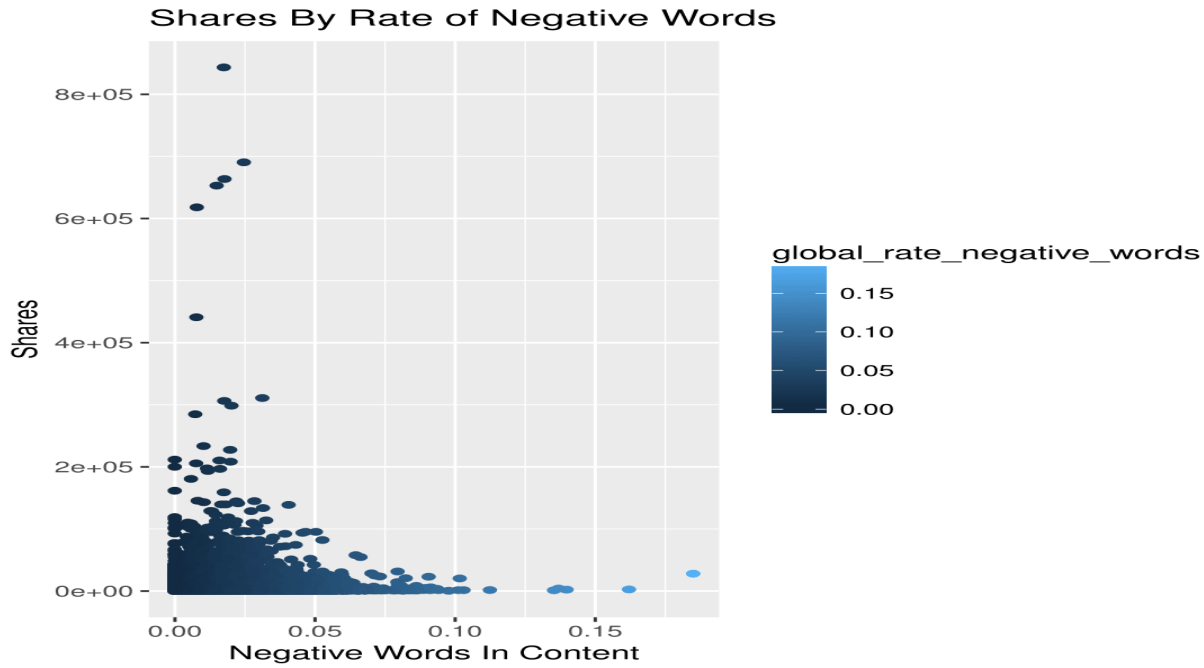
Shares by number of images:
Articles with fewer images get shared more.



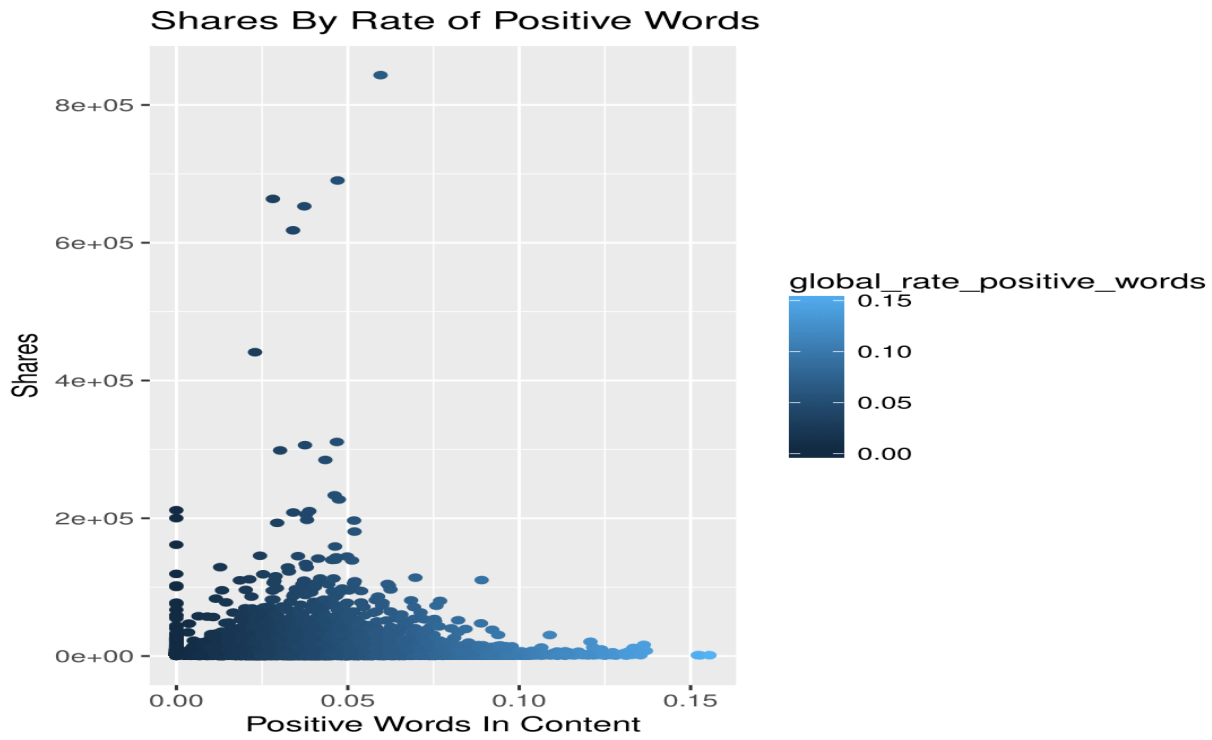
Shares by number of keywords:
Using more key words will increase the number of shares, as there or more words that can be found in search engines.



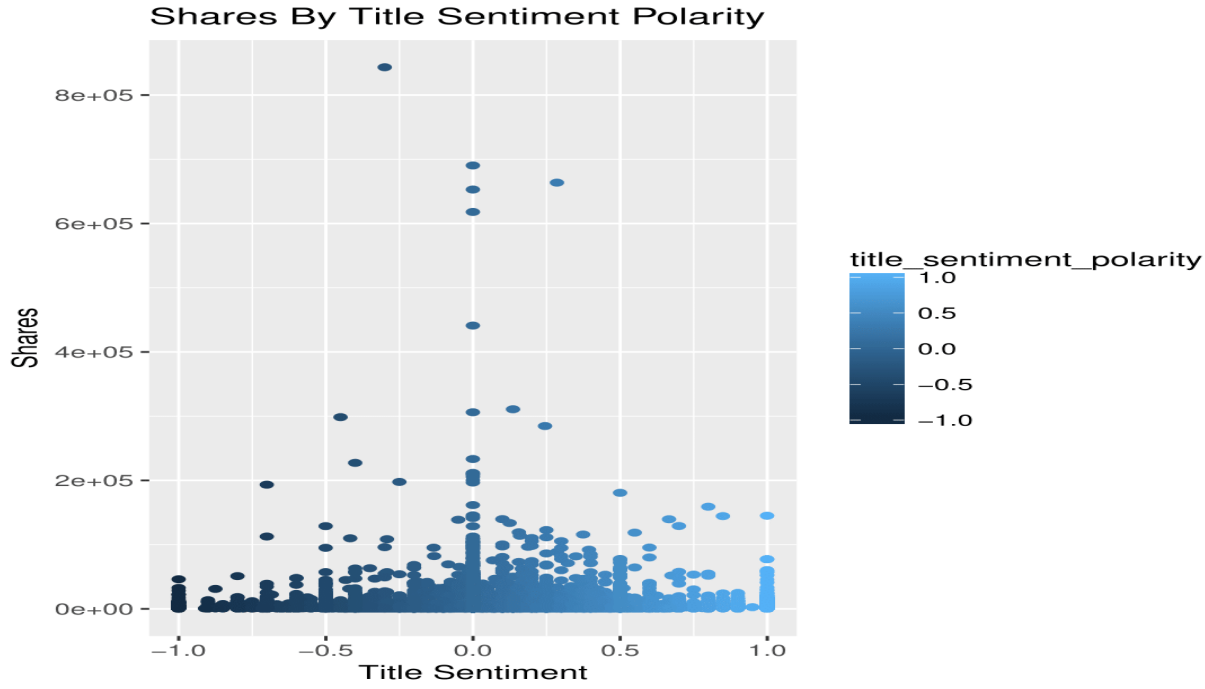
Shares by rate of negative words*:*
Using negative words will decrease the number of shares.



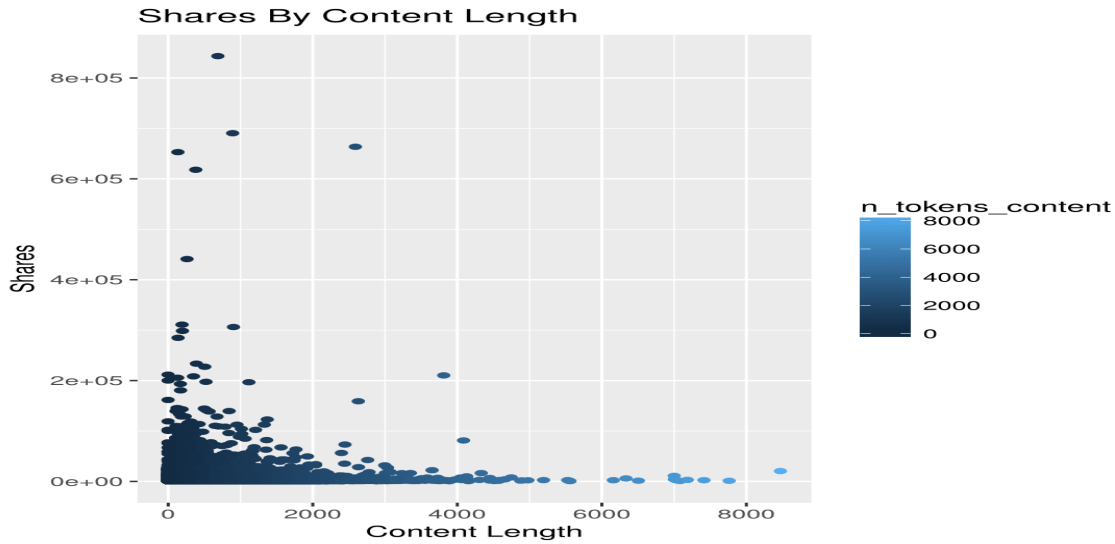
Shares by rate of positive words.
Having a global rate of positive words around 0.05, will result in the highest number of shares.



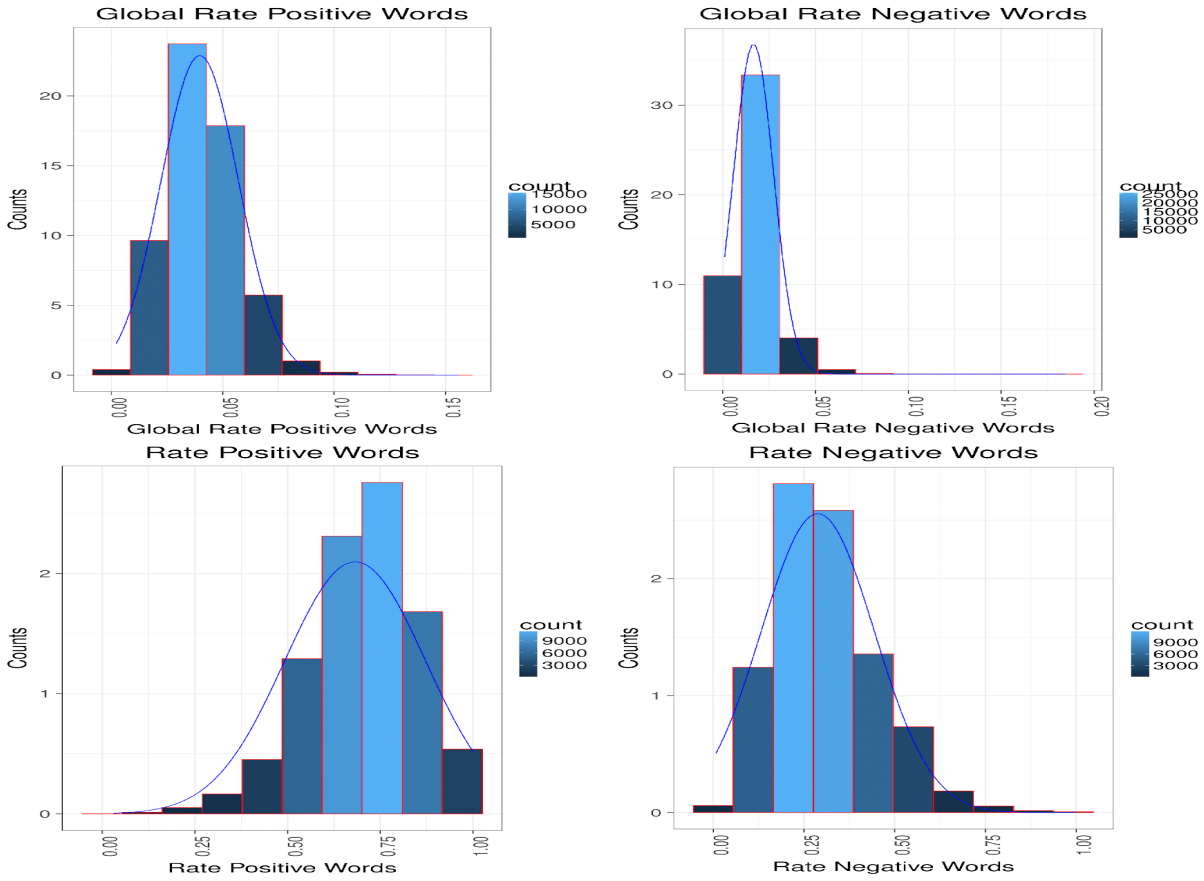
Shares by Title by sentiment polarity:
Neutral articles will produce the highest number of shares.



Shares by content length.
Shorter articles get more shares.



Shares by rate of positive/negative words:
There are higher number of articles with more positive words than negative.



4. Data preparation and cleaning

The Online News Popularity dataset csv file was not clean. The dataset has 61 attributes describing each aspect of the articles and 39,797 number of observations. The number of attributes were reduced after reading the csv file into RStudio. We removed the URL and Timedelta attributes because they were non-predictive variables. Furthermore, we removed n_token_content because it contained misleading values.

We identified extreme outliers from the numeric variables, based on the difference of the values from the central point. We imputed with the median value.

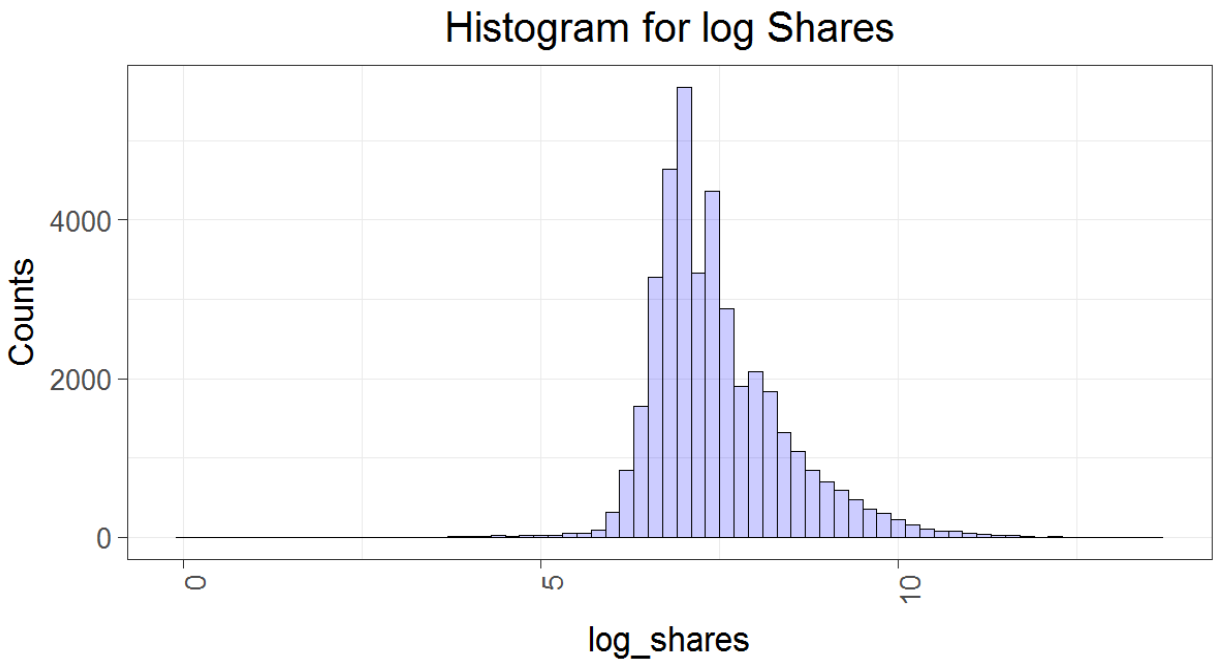
5. Data Modelling

5.1. Clustering

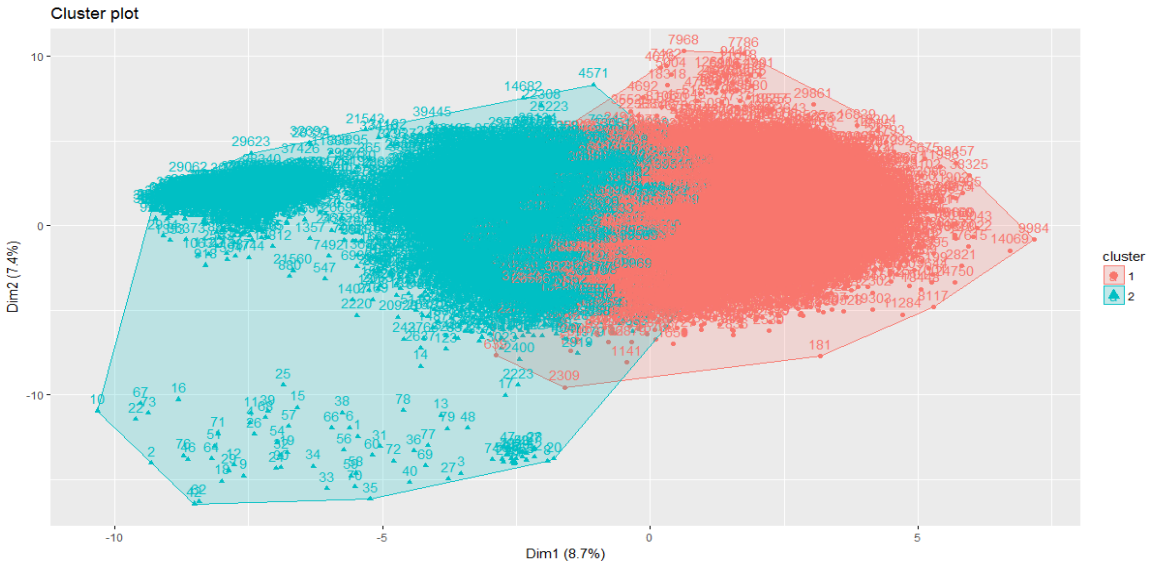
Predicting the number of shares given the independent variables is turned into predicting whether news articles are popular or not. Hence, we run into a classification task that takes as input a news article with its features and provides whether the article is popular or not.

Based on the online news popularity dataset, below the plot for the shares attribute. Based on the log shares, we consider mean+1*SD=8.405341 as threshold for labeling the news instances as follows:

`If log( shares)>  mean+1*SD, then` **`popular`** `else` **`non-popular`**



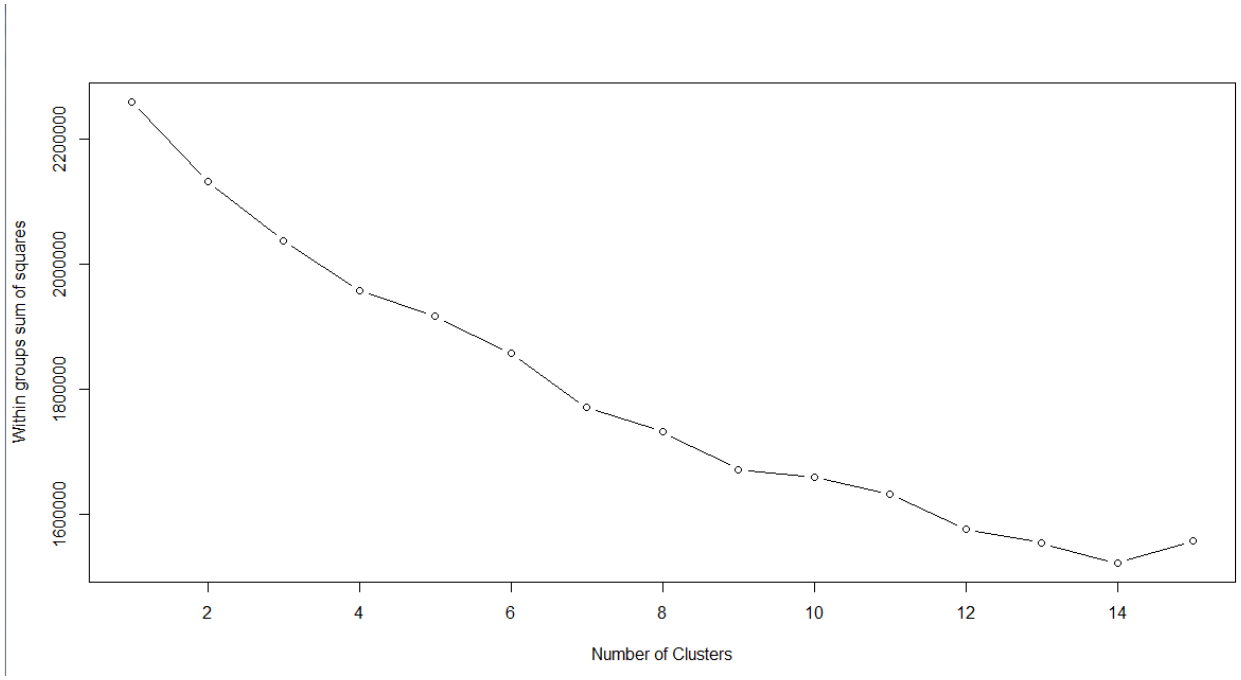
Clustering using all original attributes



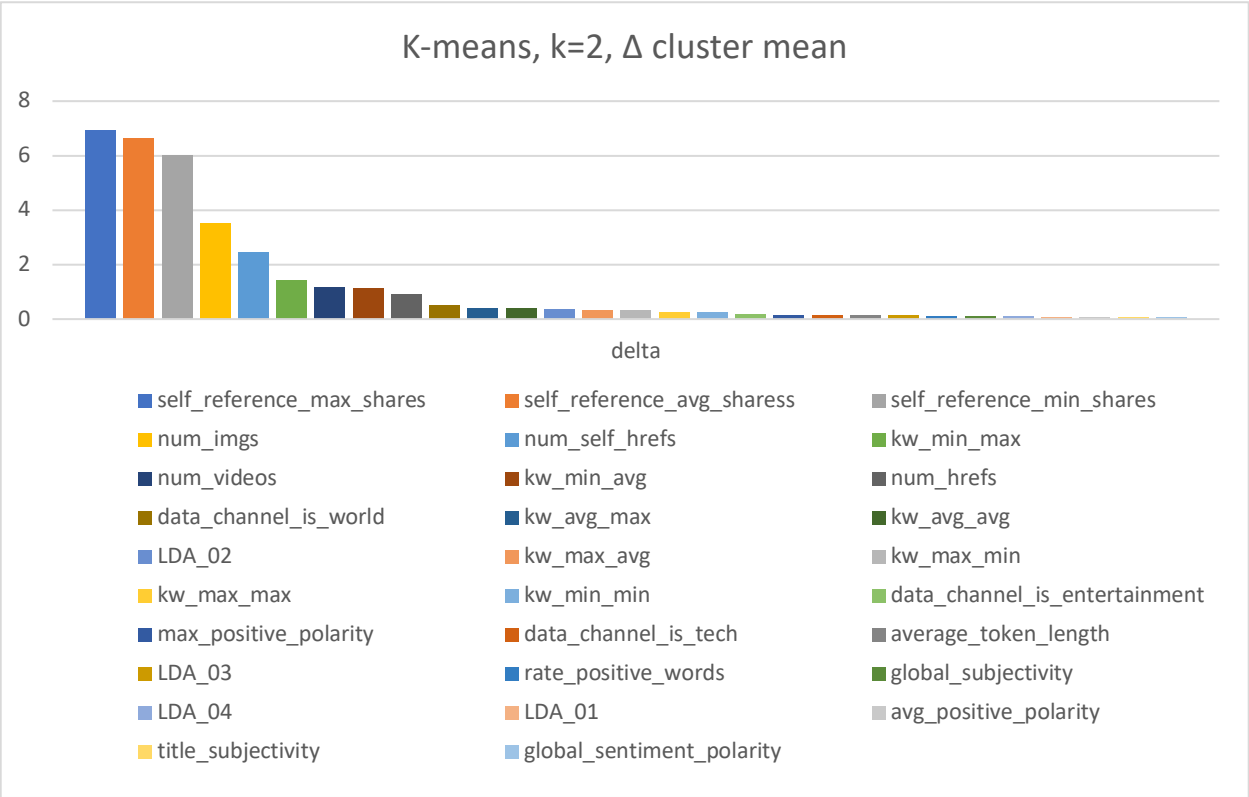
1382454.6 750143.6
(between_SS / total_SS = 5.6 %)

unpopular popular
1 22034 4525
2 11751 1334

Principal component analysis



Clustering with k=2



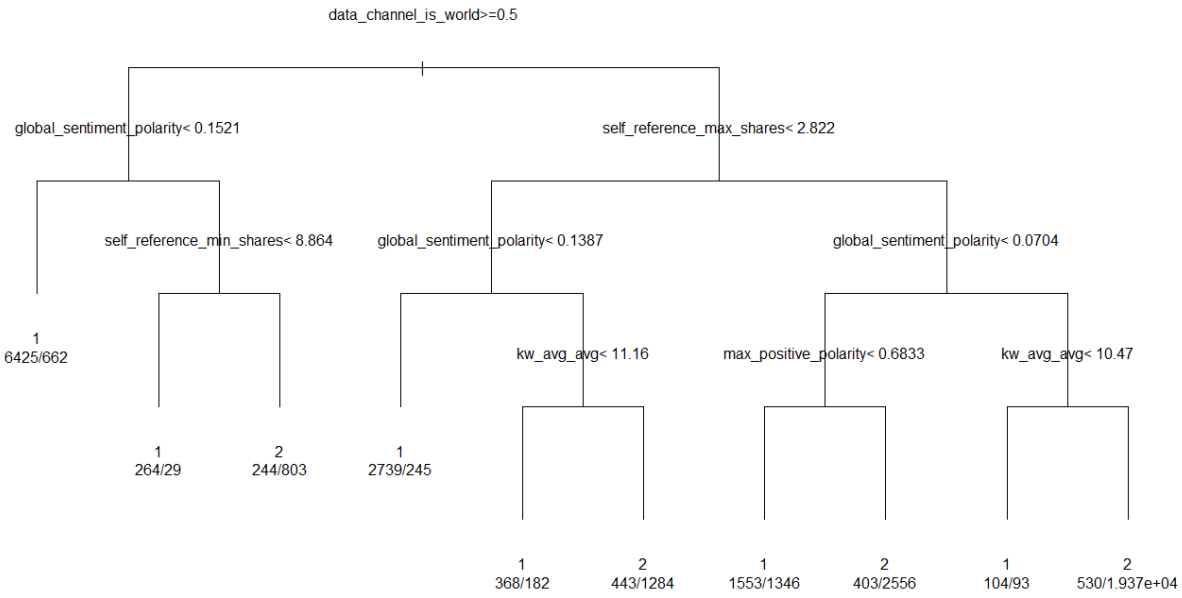


Fig: Decision tree for classifying news articles into cluster1/cluster2

Observations

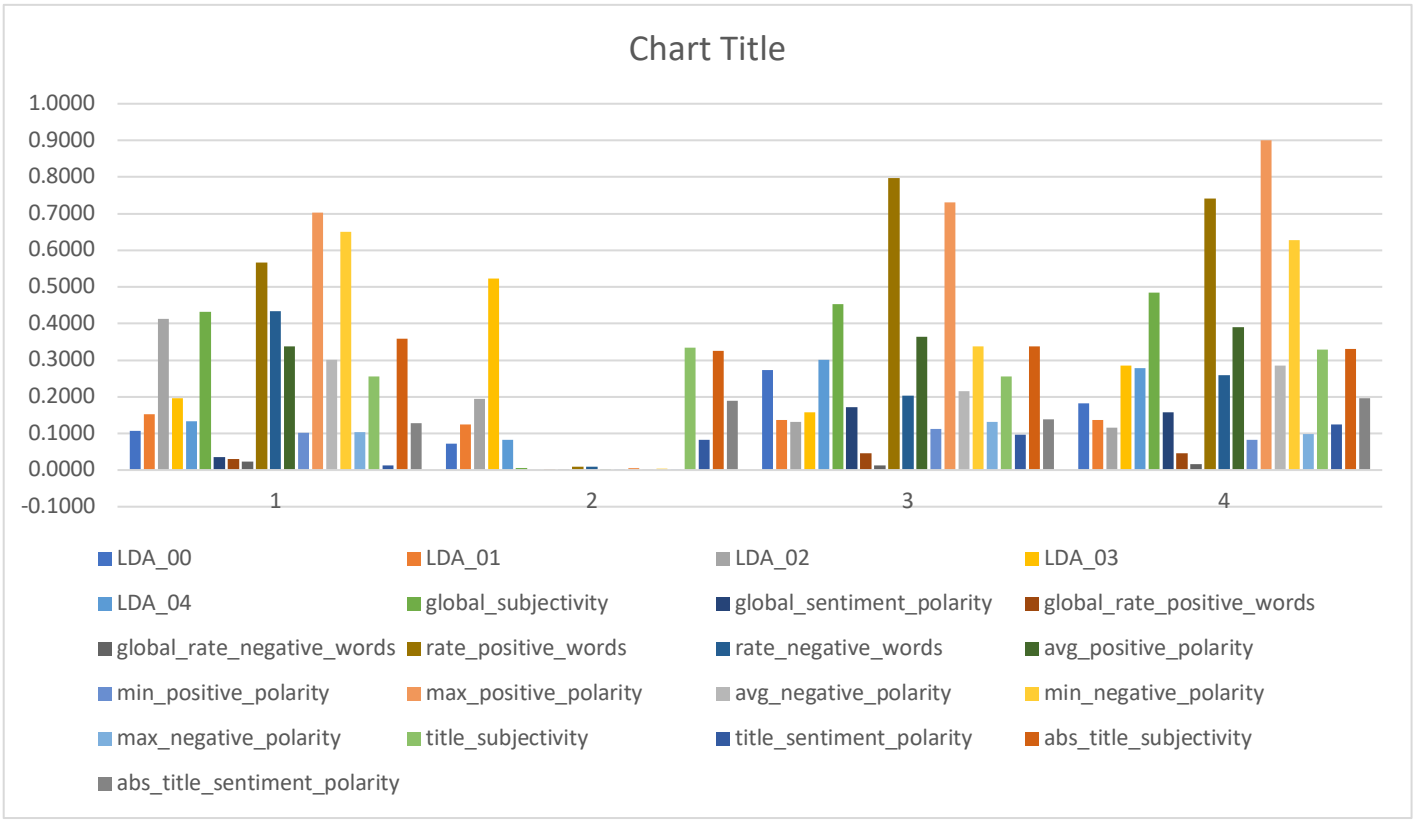
Observation 1: A major observation on the main features used in the decision tree for classifying news instances is the tendency to split based on subjective/ non subjective content. Subjectivity is when text is an explanatory article which must be analyzed in context. Polarity, also known as orientation is the emotion expressed in the sentence. It can be positive, negative or neutral.

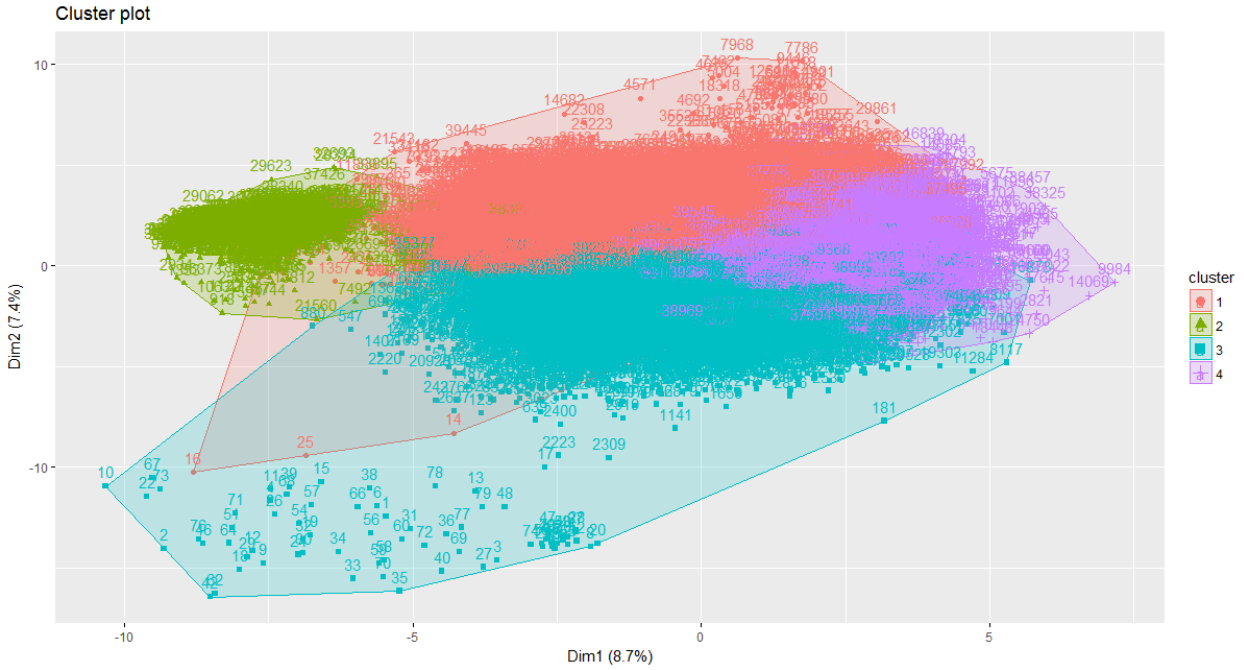
Observation 2: the root node is “data channel is world”. By using this tree, any news article that falls into any other category

Clustering with k=4

	unpopular	popular
1	11074	1309
2	943	267
3	11154	1581
4	10614	2702

Within cluster sum of squares by cluster:
[1] 111456.020 4022.218 167906.997 92283.543
(between_SS / total_SS = 41.3 %)





There is a mix of different groups of attributes that can lead to generate clusters with many overlapping characteristics with few other unique characteristics in each.

- **Subjective/unsubjective**: Clustering using k=4, separates subjective from non-subjective news. We can definitely assume cluster 1, 3, 4 have news articles that are subjective where articles that have global positive or global negative polarity are found in the same cluster.
- **Positive/negative polarity**: Cluster 3 and 4 group more positive news than negative as opposed to cluster 1 where we see balance rate of positive words compared to rate of negative words.
- **LDA topics:** we observe the following:
    - Cluster 1 has closeness to LDA002
    - Cluster 2 has closeness to LDA003
    - Cluster 3 has closeness to LDA00 and LDA004
    - Cluster4 has closeness to LDA003 and LDA004

Conclusion

The dataset is an example of vague clustering where clusters are hard to identify identities for. Clustering should be driven through the set of attributes we select since the beginning as we would like to target a key characteristic that separates the different clusters. For instance, given all input variables, we can expect clustering to result into clusters that have a mix of the following characteristics:

- Subjective/subjective
- Positive/negative polarity
- LDA topics
- Domains (business, tech, social media, etc.)

Hence, the selection of the attributes is critical for a good clustering. In the following we provide a systematic approach for attribute selection that serves in clustering online news articles.

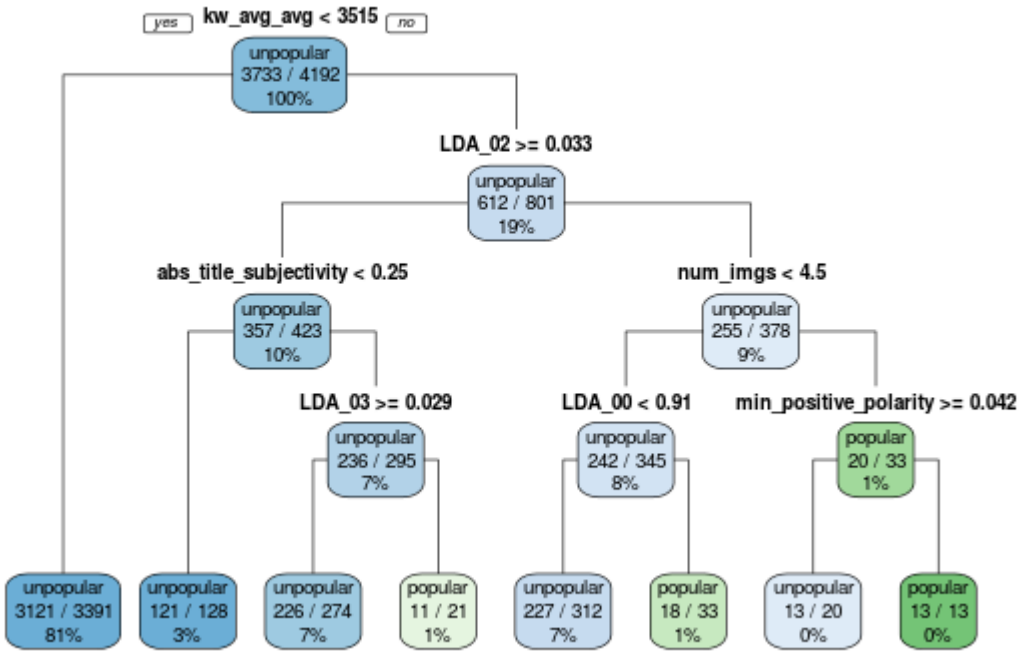
5.2. Proposal: Per-topic decision tree modelling

The solution is to find a subset of news articles that share more commonalities in terms of selected features. Subsets here can be created based on topics. There are 6 topics of news articles in the dataset: Business, technology, entertainment, social media, lifestyle and world. For each topic, a decision tree is built which enables us to generate it at a granular level. In each domain, we develop a model based on about 67% of the domain subset. The remainder, 33% of the domain subset dataset was used to test the model.

Motivation:

Classification of popular/unpopular news articles can be done through creating more refined, accurate decision trees that can better represent the features in each topic of news articles. News articles of the same category would present more important features than others.

Decision tree for business domain

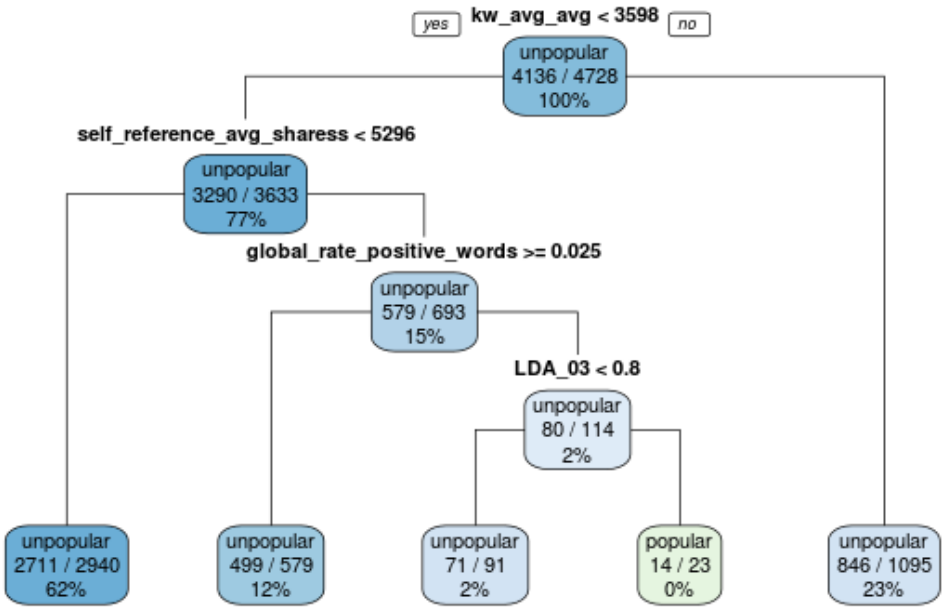


- When the average keyword has an average number of shares less than 3515, news is automatically classified as unpopular.
- If the news article has a closeness to LDA_02 larger than 0.033 (talking about LDA_02 news), and the title subjectivity is <0.25, news are automatically classified as unpopular.
- If subjectivity within LDA_02>0.033 is larger than 0.25, the classifier checks also for closeness to topic LDA_03, the news article is unpopular if LDA_03>0.029 and popular otherwise.
- Popular news is also found when num_imgs>4.5 and min_positive polarity<0.042.

Observation:

- The most important attributes present in a business decision trees are related to Natural Language Processing attributes and LDA topics.

Decision tree for entertainment domain

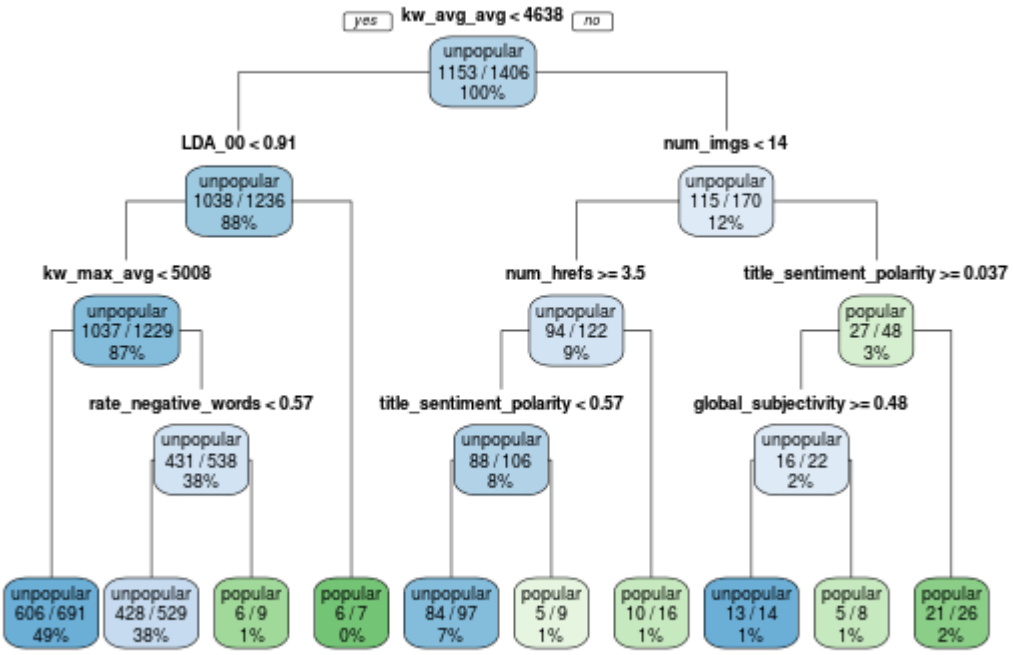


- If the average keyword received an average shares >3598, the class is **unpopular**.
- If the average keyword received an average shares <3598 and self-reference average shares < 5296 then class is **unpopular**.
- If the average keyword received an average shares <3598 and self-reference average shares < 5296 and a global rate of positive words >0.025, the class is **unpopular**.
- If the average keyword received an average shares <3598 and self-reference average shares < 5296 and the global rate of positive words <0.025, and closeness to LDA_03 >0.8, the class is **popular**.

Observation:

- The most important attributes present in **entertainment** decision trees are related to Natural Language Processing attributes, topics and shares attributes.

Decision tree for lifestyle domain

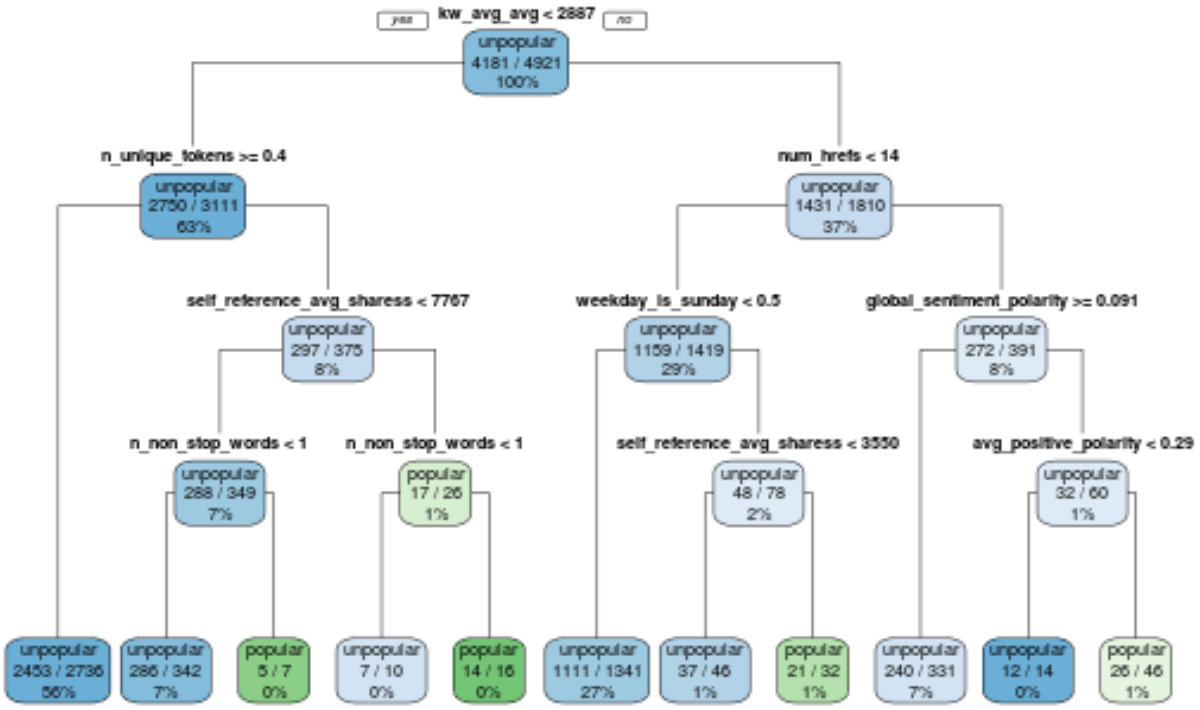


- If the average keyword received an average shares <4638, and the topic of the article is LDA_00 with high closeness (>0.91) class is **popular**.
- If the average keyword receives an average shares <4638 and the topic of the article is LDA_00 with closeness (<0.91) and the average keyword in the article received a max shares >5008 and rate negative words >0.57 class is **popular**.
- If the average keyword receives an average shares >4638 and the number of images >14, and the article is neutral (title sentiment polarity <0.037) class is **popular**.

Observation:

- The most important attributes present in **lifestyle** decision trees are related to Natural Language Processing attributes, digital attributes (num images) and topics.

Decision tree for technology domain

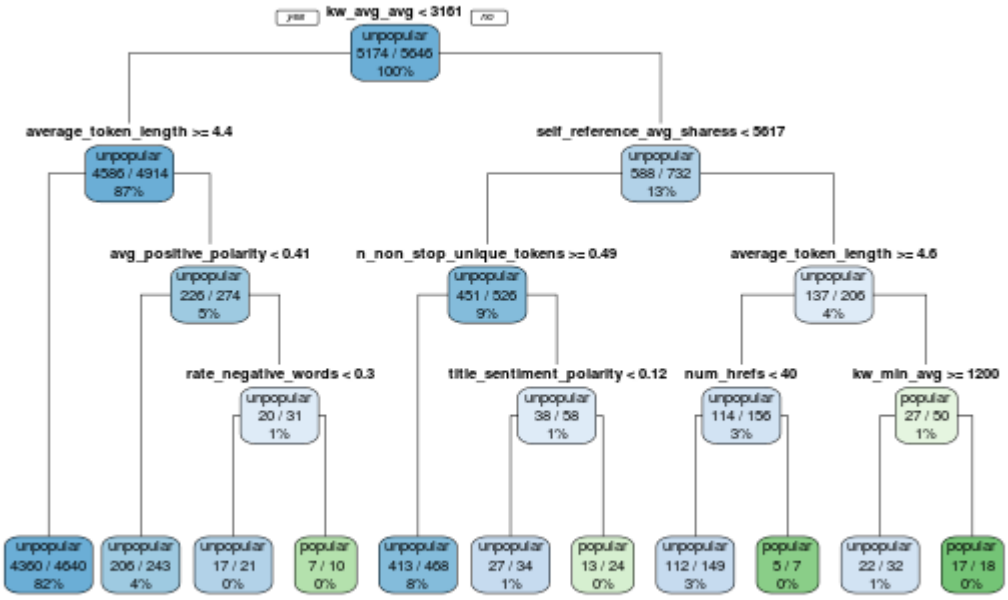


- The articles with num of links <14, and published on Sunday, and the self-reference average shares is >3550, class is **popular**.
- The articles with num of links <14, and global sentiment polarity <0.091, and the average positive polarity is >0.29, class is **popular**.

Observation:

- The most important attributes present in **technology** decision trees are related to shares attributes, publication days, and Natural Language Processing attributes.

Decision tree for World domain



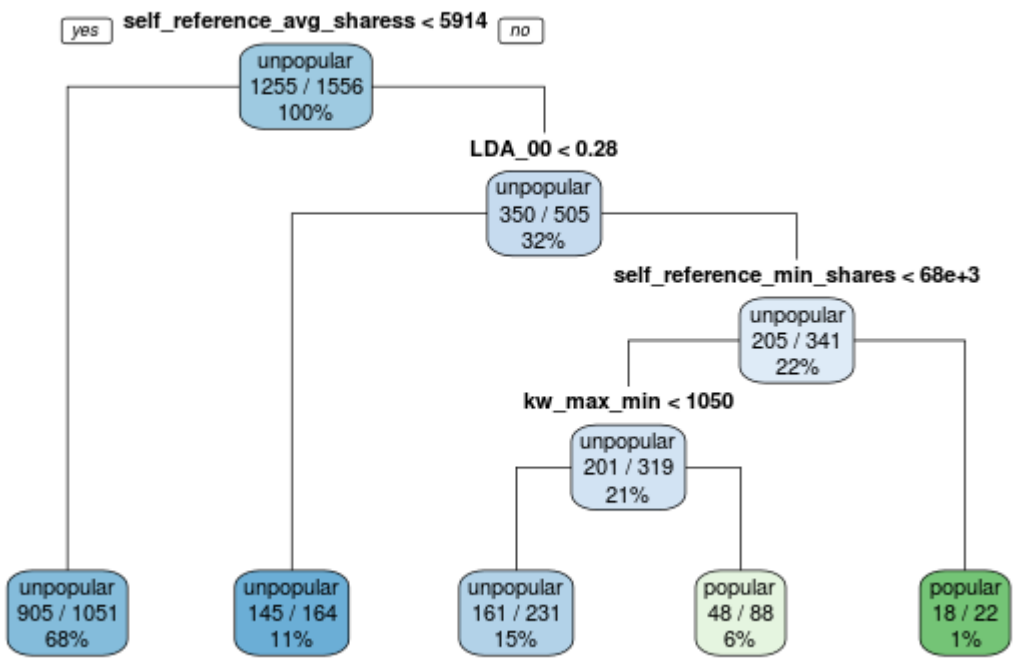
We believe the quality of this decision tree is questionable as we see some keywords attributes such as average token length, number of non-stop unique words are used in the top levels of the tree for splitting. We ignore these attributes in our explanation.

- If average positive polarity >0.41 and rate negative words >0.3 then class is **popular**. This could be related to articles that have both positive and negative emotions which tend to make them controversial and popular.

Observation:

- The most important attributes present in world decision trees are related to word attributes (tokens, keywords, etc..) and Natural Language Processing attributes.

Decision tree for the social media domain



- If self-reference average share >5914, and closeness to LDA_00 > 0.28 and self-reference minimum share >68000 then class is **popular**.
- If self-reference average share >5914, and closeness to LDA_00 > 0.28 and self-reference-minimum share<68000 and the worst keyword receives max shares of >1050, then the class is **popular**.

Observations:

- The most important attributes present in social media decision trees are related to shares attributes (19 to 30 based on UCI web site) and LDA topics.

Classification accuracy

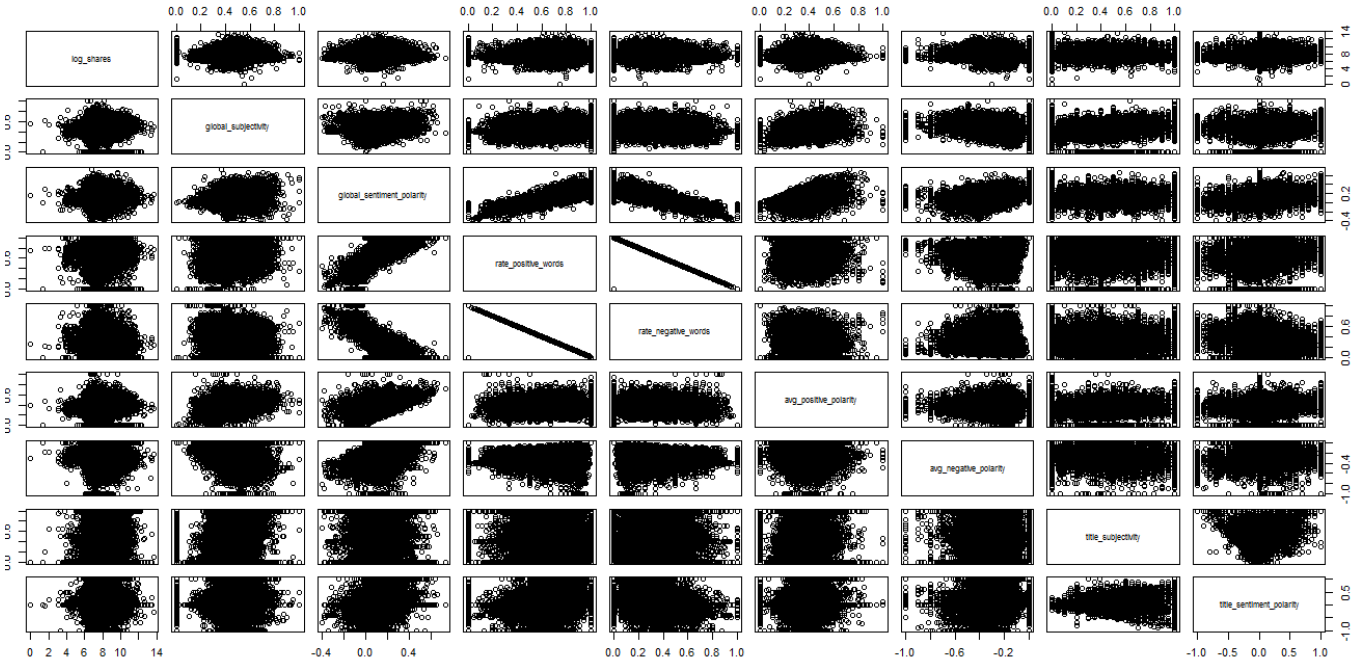
On each domain subset, the corresponding decision tree classifier is tested on the testing data. The best classifier that gives the highest accuracy is naïve Bayes.

Overall accuracy			
Domain	DT	RF	NB
Business	88%	88%	**92%**
Lifestyle	82%	81%	**93%**
Entertainment	88%	87%	**91%**
Tech	85%	85%	**87%**
World	91%	91%	**92%**
Social Media	76%	78%	**92%**

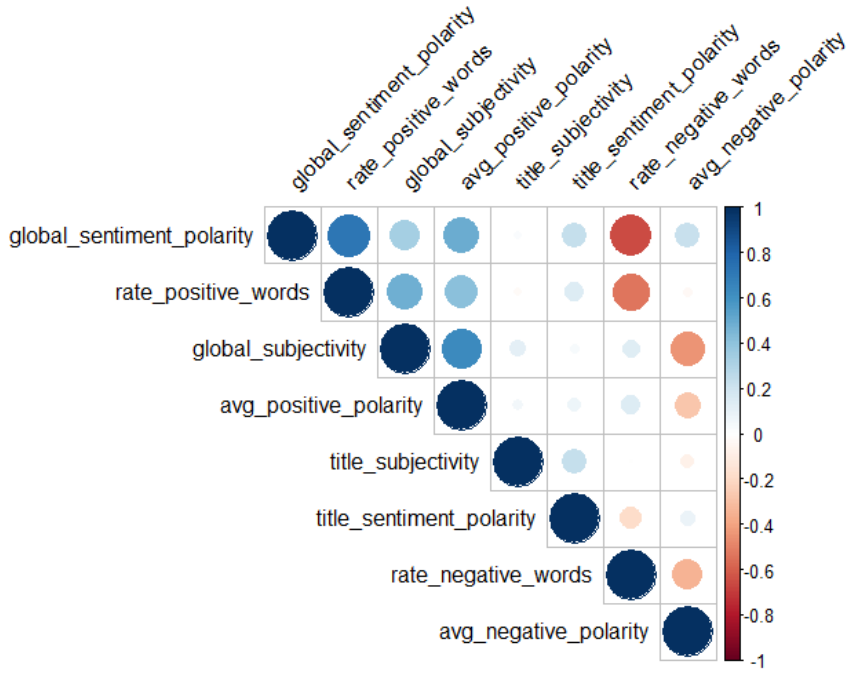
The popular/unpopular accuracy is given in the table below. Naïve Bayes gave the best performance in predicting popular news. Predicting popular news is the most accurate on the business domain (70%) as compared to other domains.

Popular class accuracy			
Domain	DT	RF	NB
Business	0.60 %	0.30%	**70.20%**
Lifestyle	2.00 %	0.60%	**15.00%**
Entertainment	0.90 %	0.40%	**7.00%**
Tech	0.30 %	0.50%	**14.00%**
World	0.30 %	0.10%	**7.00%**
Social Media	3.00 %	2.00%	**14.00%**

The main reason is that Naïve Bayes is based on the independence assumption between attributes and in the dataset, as per the explorative data analysis, it was found that there are few correlations between the attributes. The figure below shows the scatterplots between shares and NLP attributes. Clearly, there a linear relationship between global sentiment polarity, rate positive words, and rate negative words.



The following figure shows the correlation values between the NLP attributes. The above mentioned linear relationships show that the involved attributes are highly correlated.



5.3. Conclusions and recommendation

Our main conclusion in regards to this dataset is that the selection of attributes is critical for applying clustering and classification of popular/unpopular news articles. For this purpose, we present in the following a thorough investigation of the features for attribute selection.

Examining topics attributes

LDA_00, LDA_01, LDA_02, LDA_03, LDA_04 versus attributes 13-18

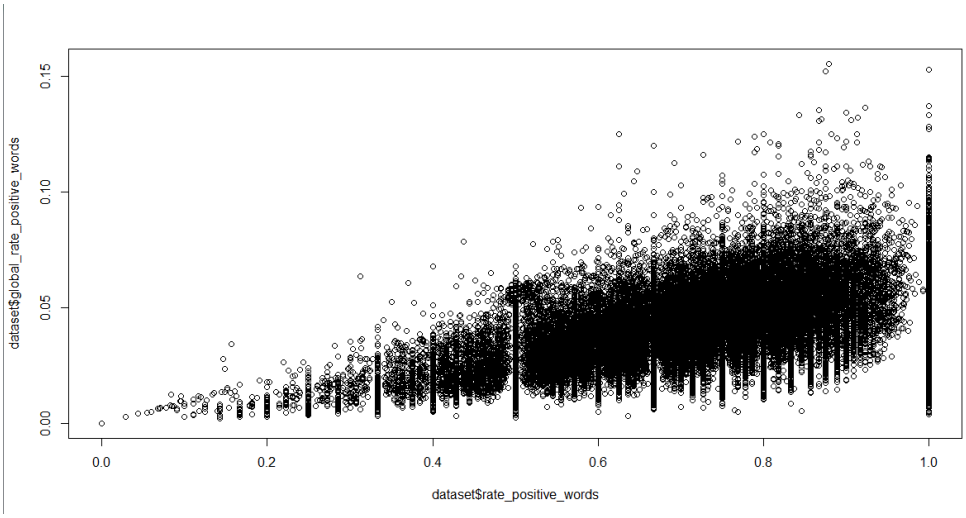
News articles are already manually categorized into business, tech, lifestyle, entertainment, socmedia and world. These are represented with binary attributes 13-18. On the other hand, LDA algorithm was applied on the whole dataset to find the most important topics, where each news article is checked for closeness against each topic. Using LDA00 to LDA04 and also attributes 13-18 will mislead the clustering process as the attributes from the two categorization (LDA and manual) of articles interfere but there are treated as independent by the clustering process. There is no way for the clustering to find relationships between these attributes. Hence we decided to remove attributes 13-18.

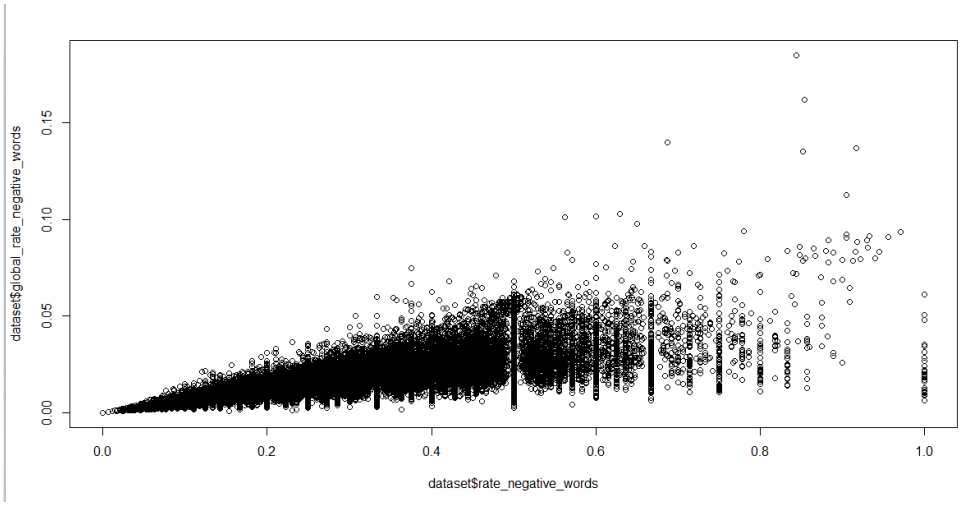
Examining NLP attributes

- Negative/positive polarity:
  A news article has Min_negative_polarity, Max_negative_polarity, Avg_negative_polarity. By considering the min and max negative polarity can mislead how observations can be clustered with

respect to polarity. Hence, we select to use only the Avg_negative_polarity *-1 for representing the negative polarity of a news article. The same applies for average positive polarity.

- Global rate positive/negative words:
  We decided to remove these columns as the information is already represented with rate positive words but by dividing over the length of the article. The figures below clearly show the linear relationship between the global rate positive words and positive word with correlation equal 0.6286261. The correlation between global rate negative words and rate negative words is 0.7795556.





- Abs title subjectivity/ Abs title sentiment polarity
  The dataset contains title subjectivity [-1 1] and abs title subjectivity [0 1]. We recommend to remove abs title subjectivity as it is derived from title subjectivity and does not tell negativity or positivity. The same applies on title sentiment polarity and Abs title sentiment polarity and we recommend removing Abs title sentiment polarity.

Deployment

Our project has shown how challenging it is to define what a popular news article means. he findings and the model we have developed in this project could help news media companies' managers to understand how to deliver content.

Monitoring

Regularly communicate with and solicit feedback from stakeholders.

● Ensure regular and continuous collection of news popularity data

● Compare model results with future results of news popularity data in order to validate findings and reveal new opportunities to improve the model

● Review any changes to project objectives to maintain relevance of the model.

Review

What could have been done better:

● Data cleaning: remove features as per the recommendation section.

● The modeling process: use cost matrix for boosting the prediction accuracy of popular news.