
Mashable

What is Mashable

Media and entertainment company for super fans

CHANNELS

Video

Entertainment

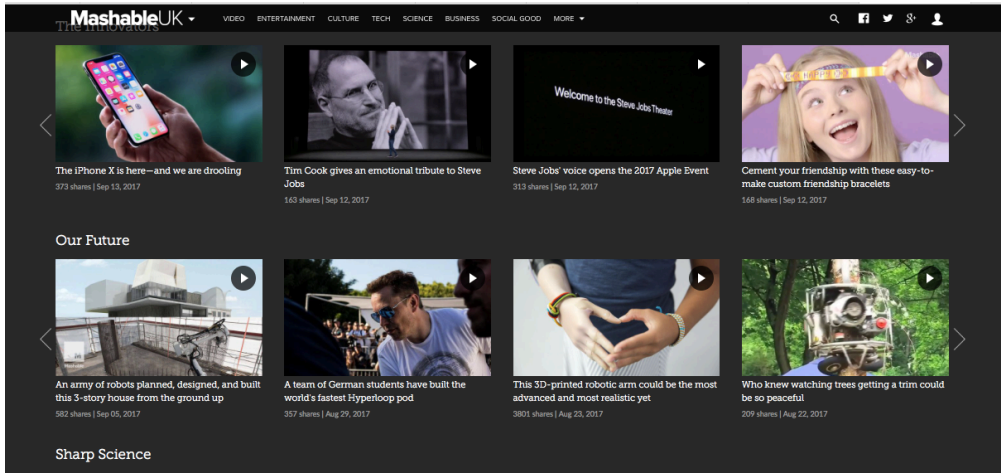
Culture

Tech

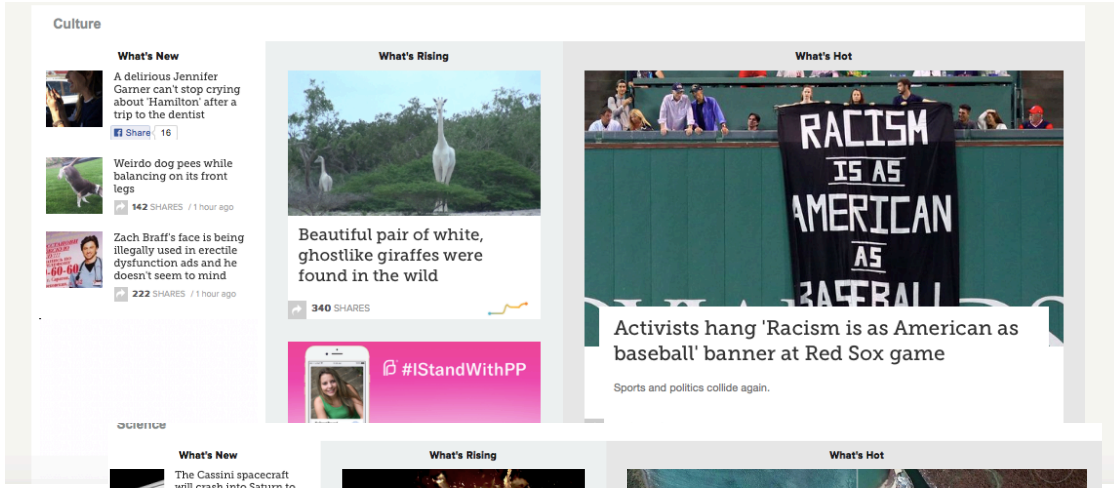
Science

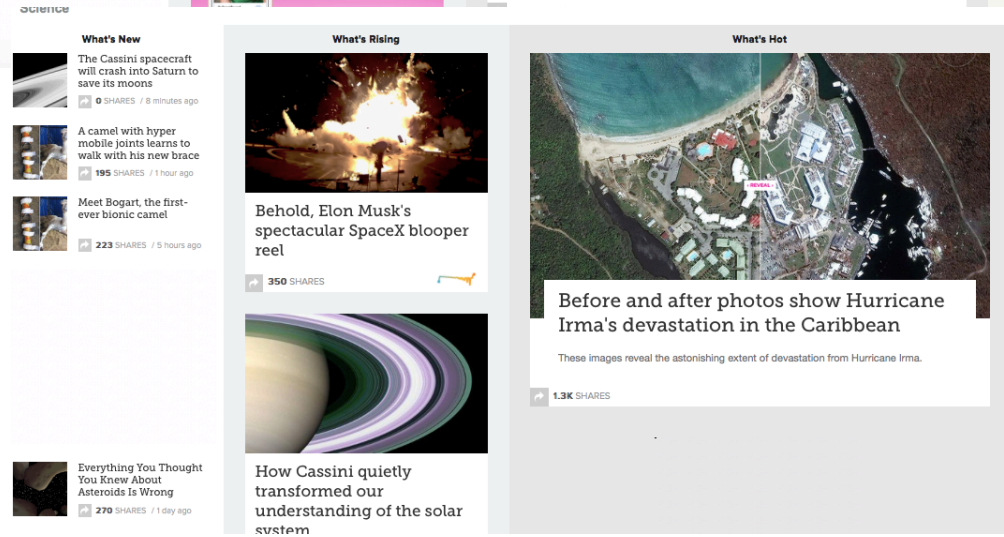
We know **the future of TV** *looks nothing like the past. Great TV won't be made for mass audiences. It'll be made for the right audiences,* ***using data*** *both* ***to*** inspire creativity and ***connect shows with influential viewers****," said Pete Cashmore, Founder and CEO of Mashable. "It won't happen on the big screen. It'll happen on the screen you have in your pocket -- the mobile phone.* The future of video is on the handset, not the TV set."

Video

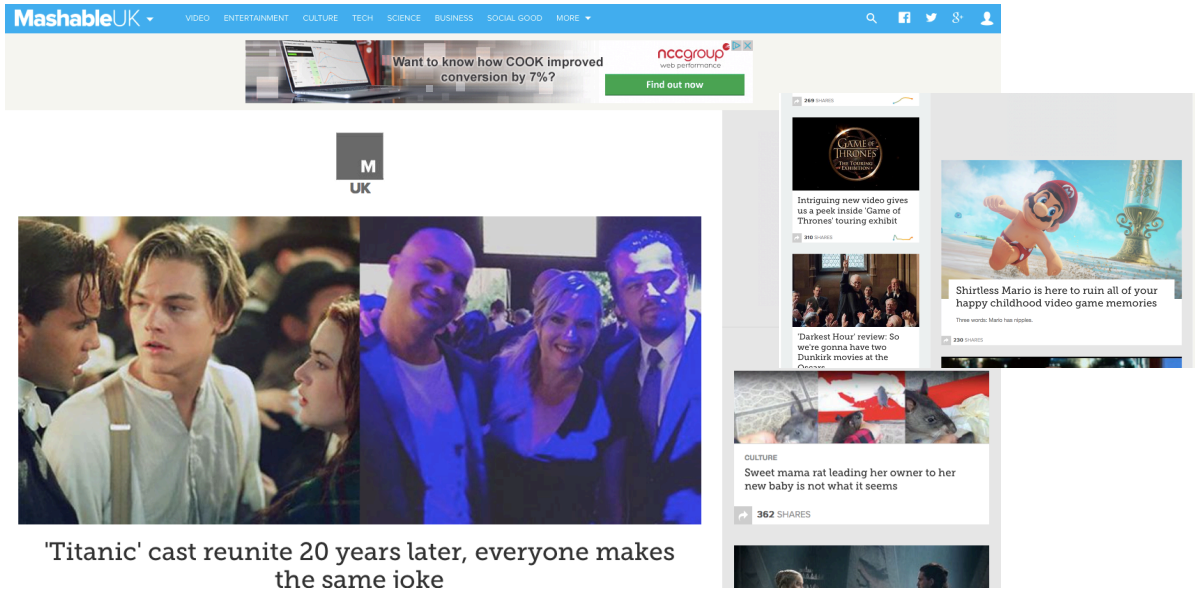


Culture





Entretenimiento



What is the goal?

PREDICTING THE POPULARITY OF ONLINE NEWS

Based on number of social shares of articles

Dataset Description

a) Data source: UCI ML Repository

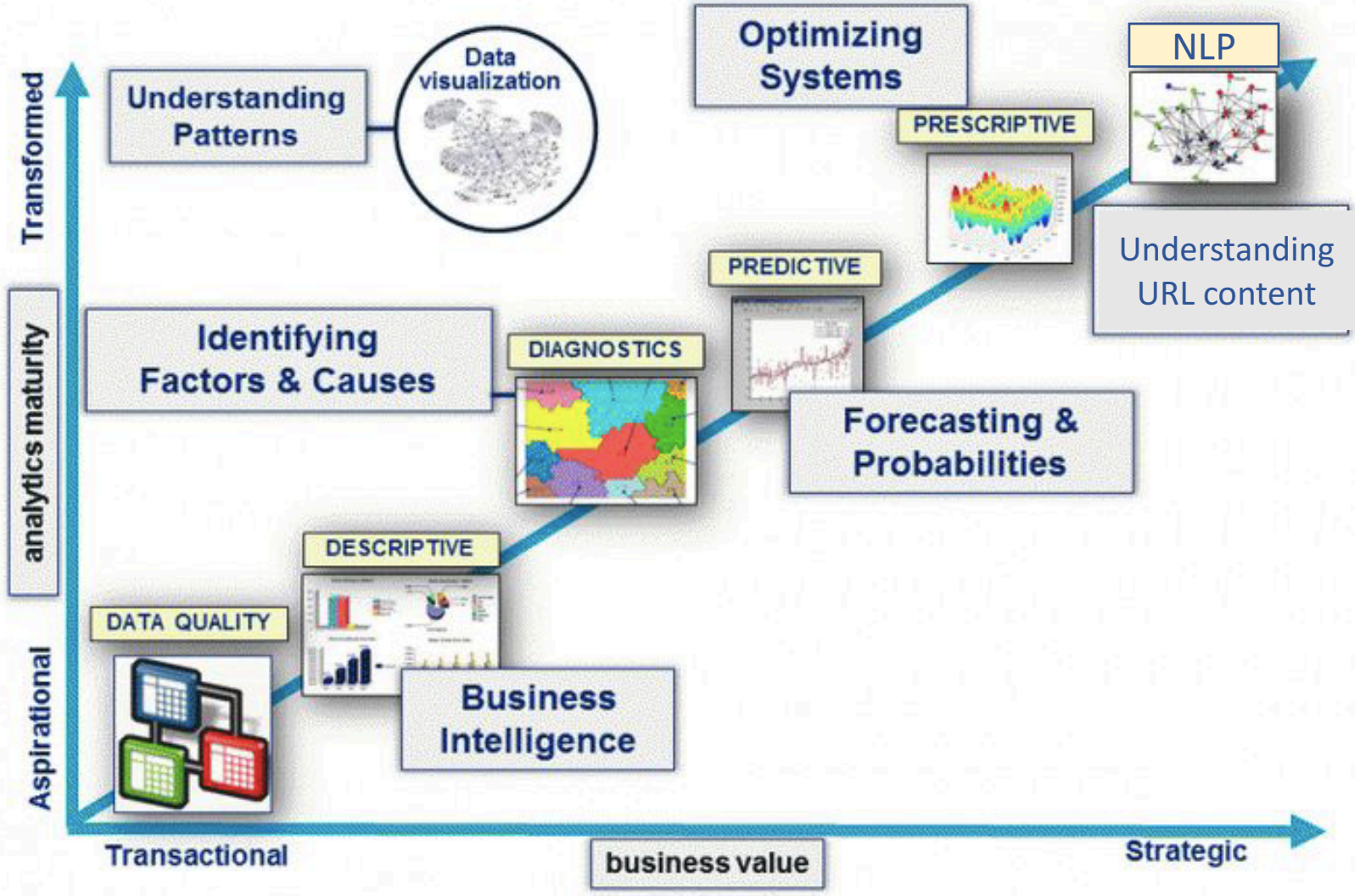
https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

- Number of Attributes 58
- Number of Records 39664

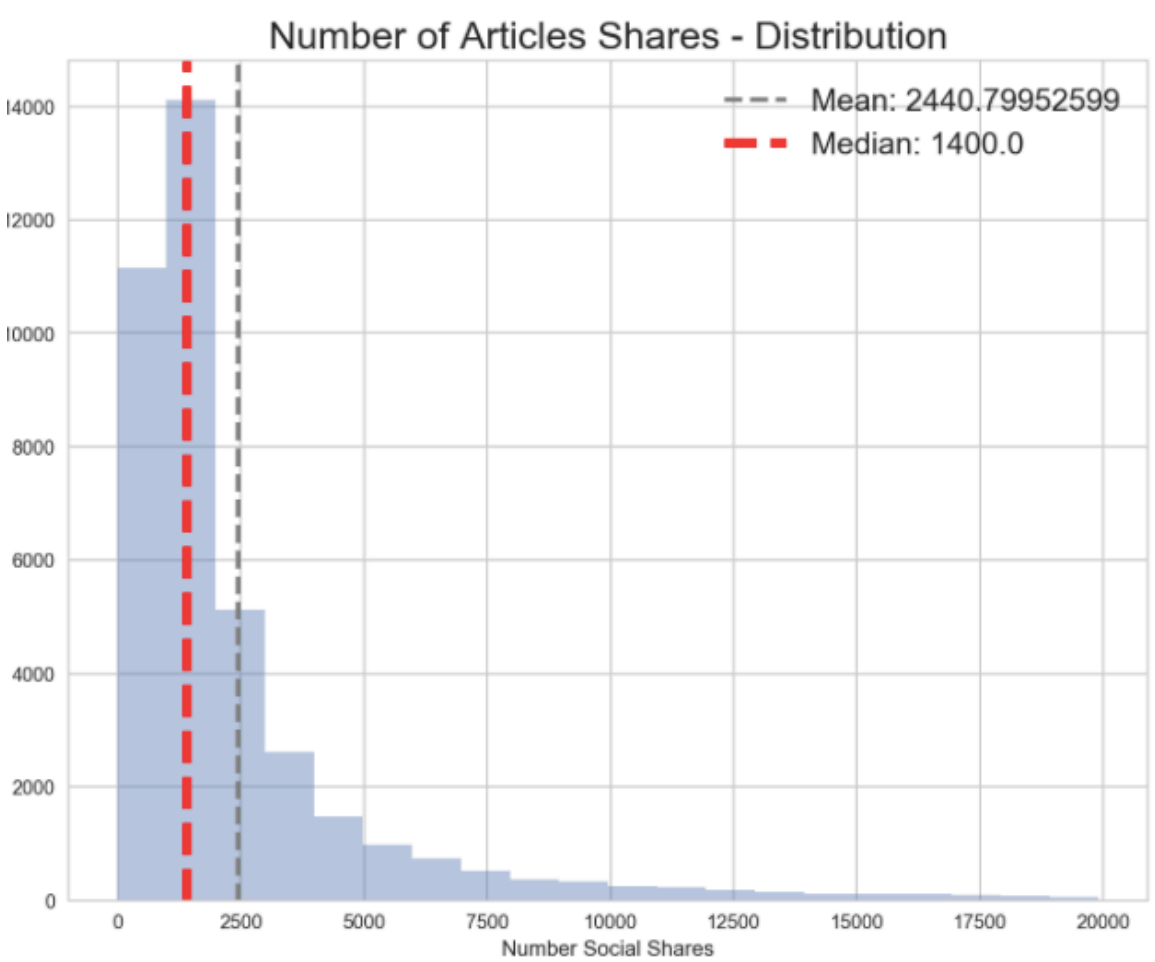
b) Web Scraping: Importing a dataset from the web and saving as a local PostgreSQL

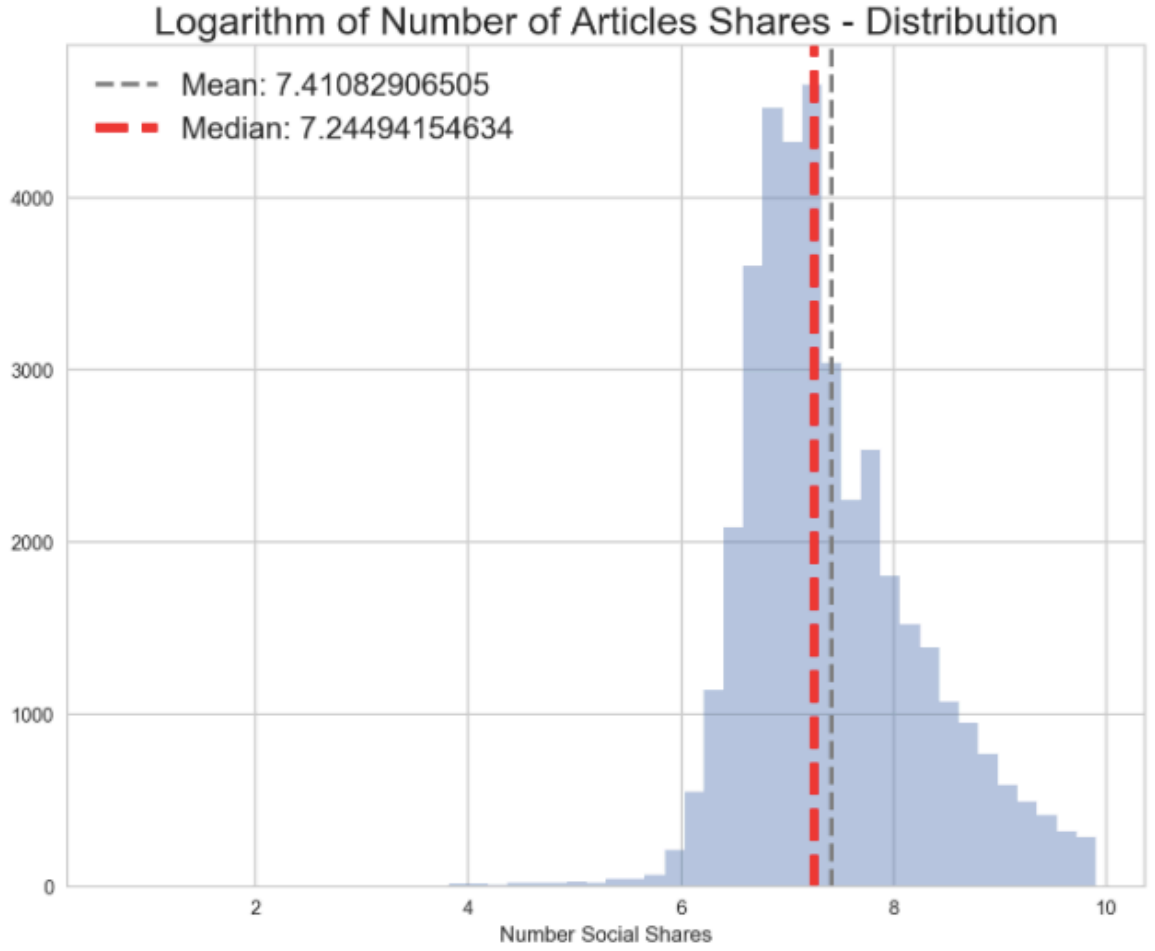
c) Importing a dataset from the URL of the article

Data Analysis



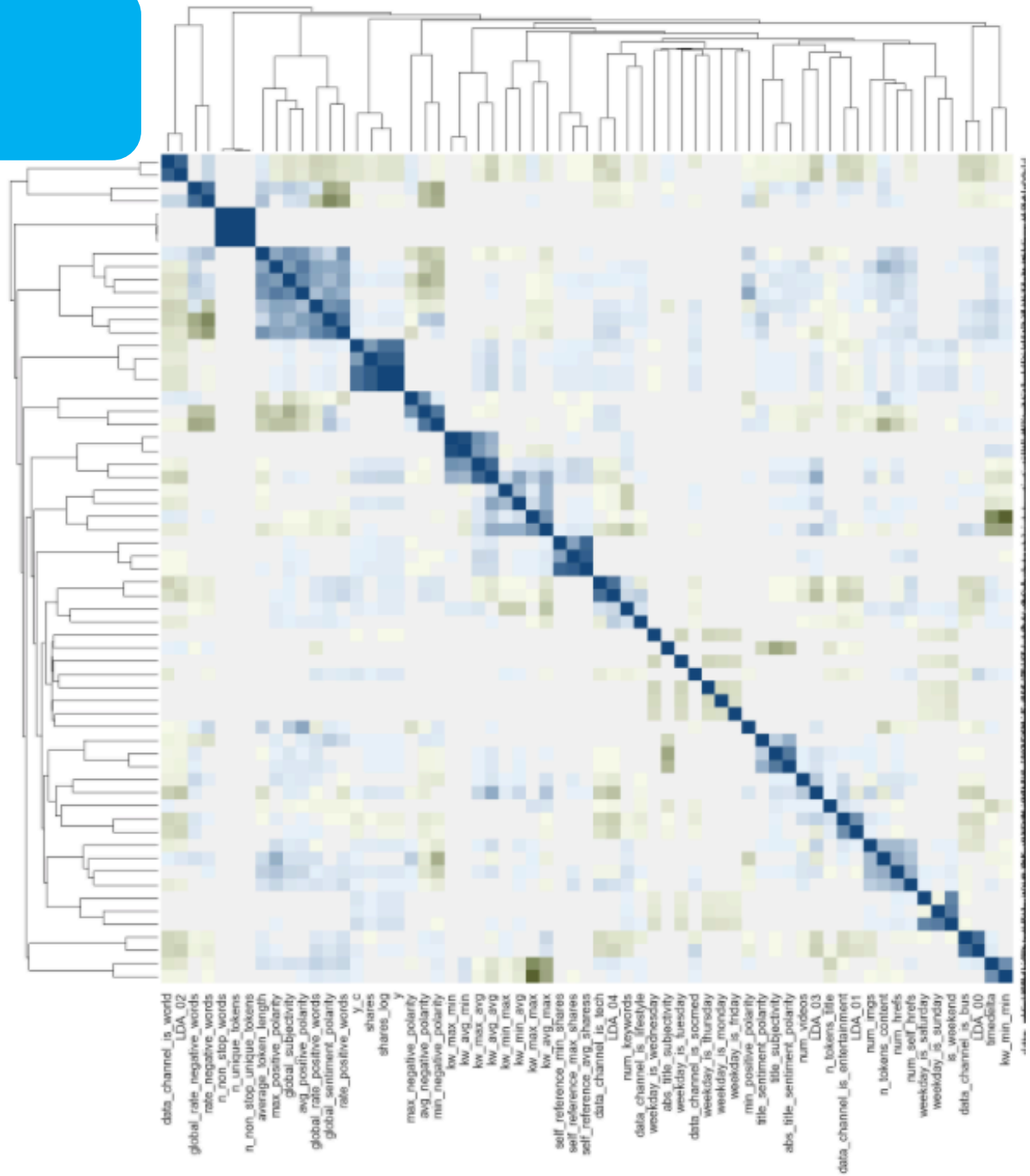
Target


Number of Articles Shares - Distribution
Mean: 2440.79952599
Median: 1400.0
14000
12000
10000
8000
6000
4000
2000
0
0
2500
5000
7500
10000
12500
15000
17500
20000
Number Social Shares

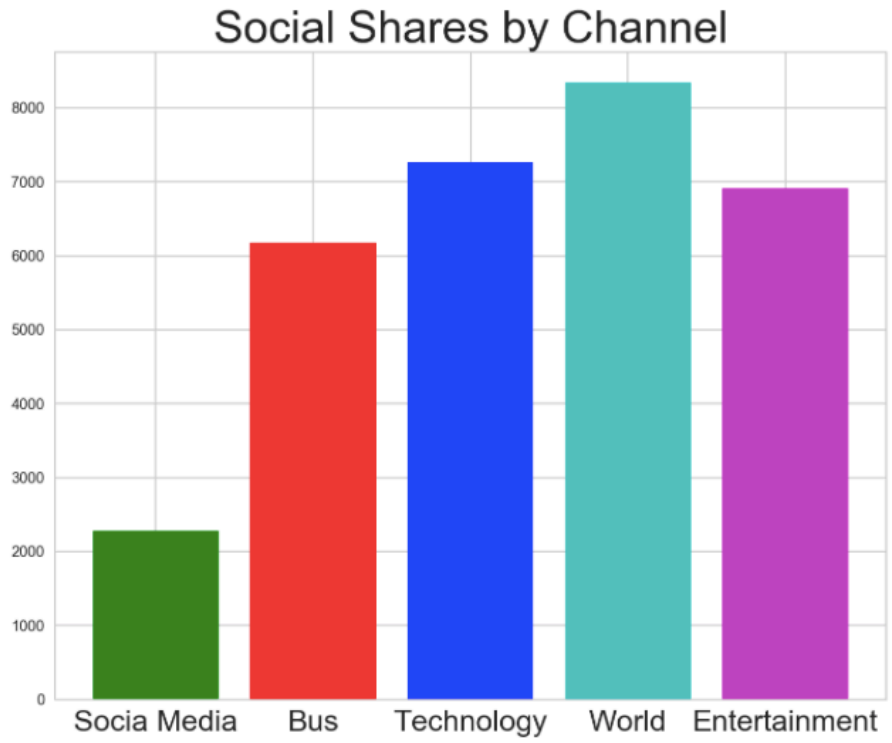

Logarithm of Number of Articles Shares - Distribution
Mean: 7.41082906505
Median: 7.24494154634
4000
3000
2000
1000
0
2
4
6
8
10
Number Social Shares

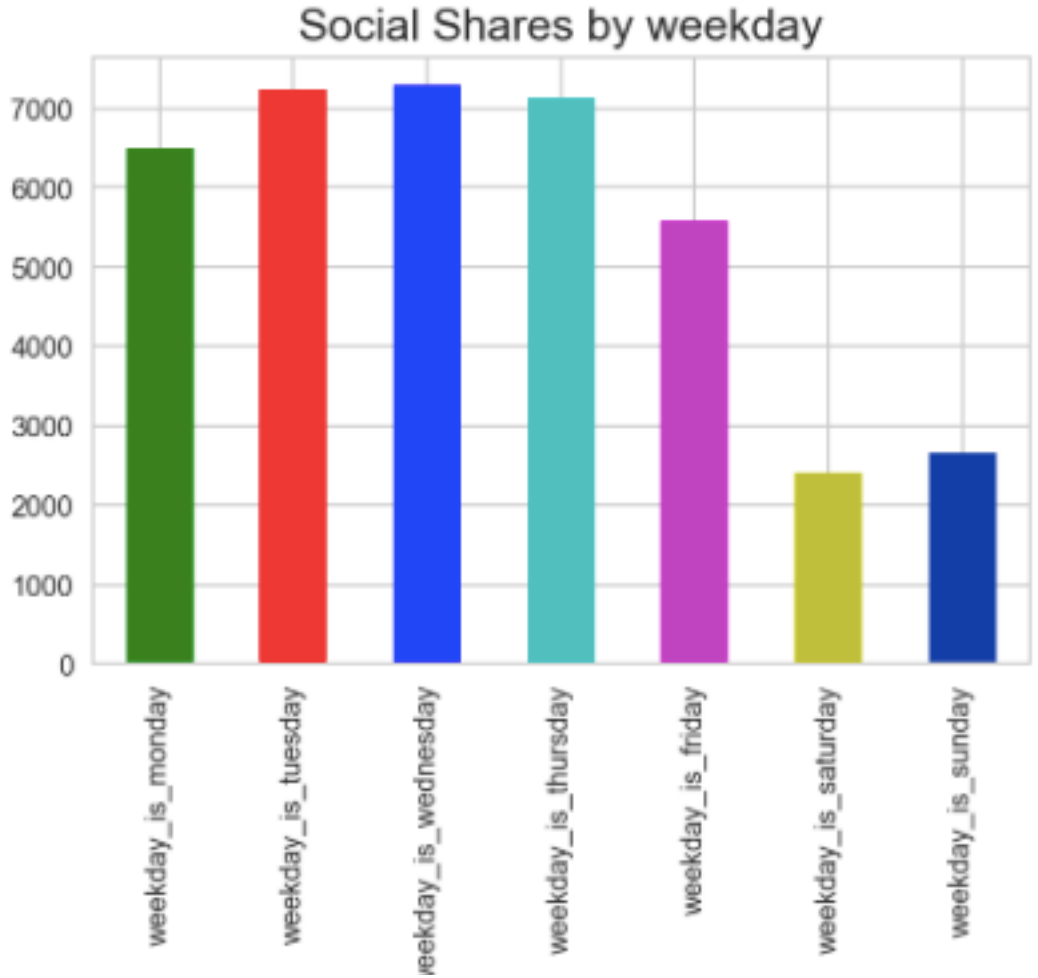
Correlation

Strong correlated features corresponding to the same topics



News in World and Technology of most popular and more shared during the middle of the week






```
# Let us see what is happening during weekends
df.groupby('is_weekend').mean().T
#                                  WEEKDAYS   vs    WEEKENDS
```

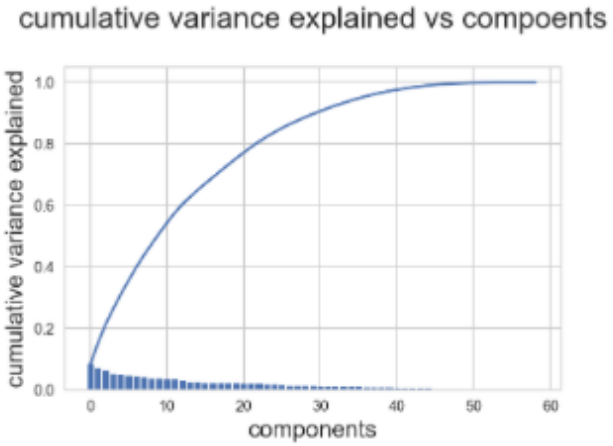
self_reference_avg_sharess	6287.758556	5998.758401

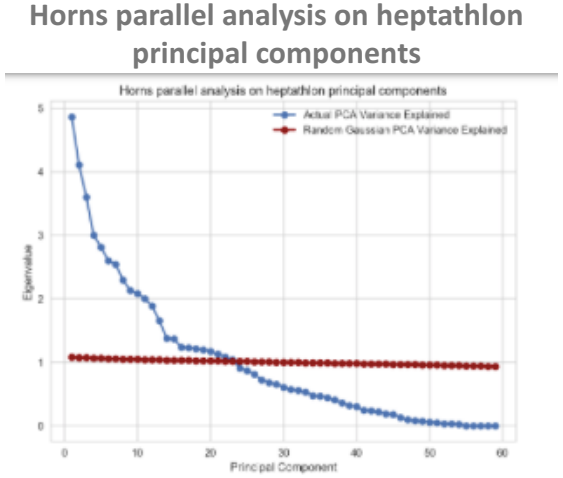
Caveats and limitations in the data set

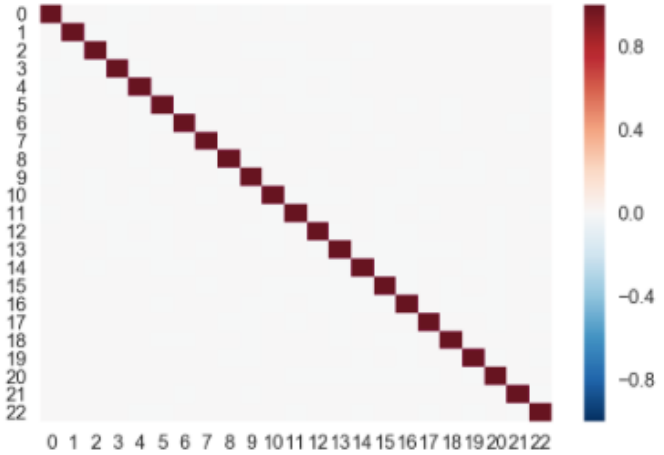
- There is no information about the relationship between the number of times an article is shared vs the amount time the article was online
- There is no information on how the channels cross over
- Criterion to self referenced articles in Mashable
- Limited information about natural language processing features

Almost 80% of the explained variance can be explained by just 23 of the principal components

Dimensionality reduction







Clustering Analysis

Number Cluster = 3

K-Means

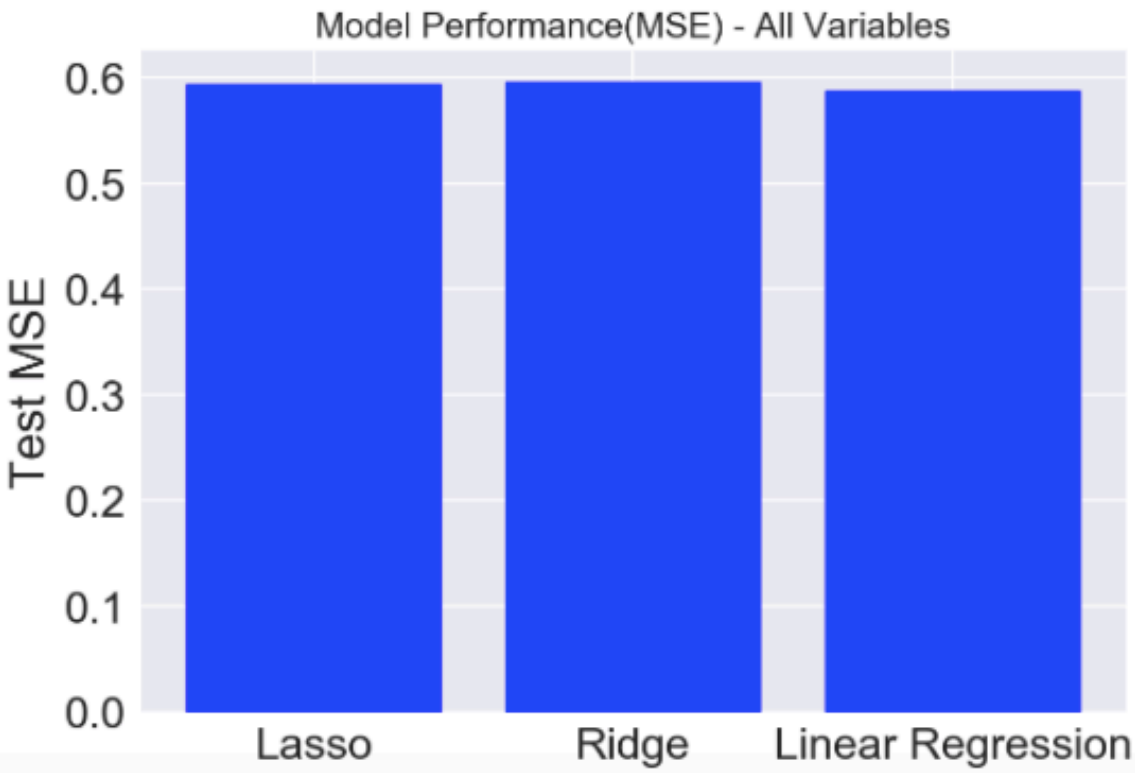
```
KMeans clustering:
Silhouette score: 0.207201788072
Homogeneity score: 0.195510046802
Completness: 0.735793987952
```

Hierchical clustering

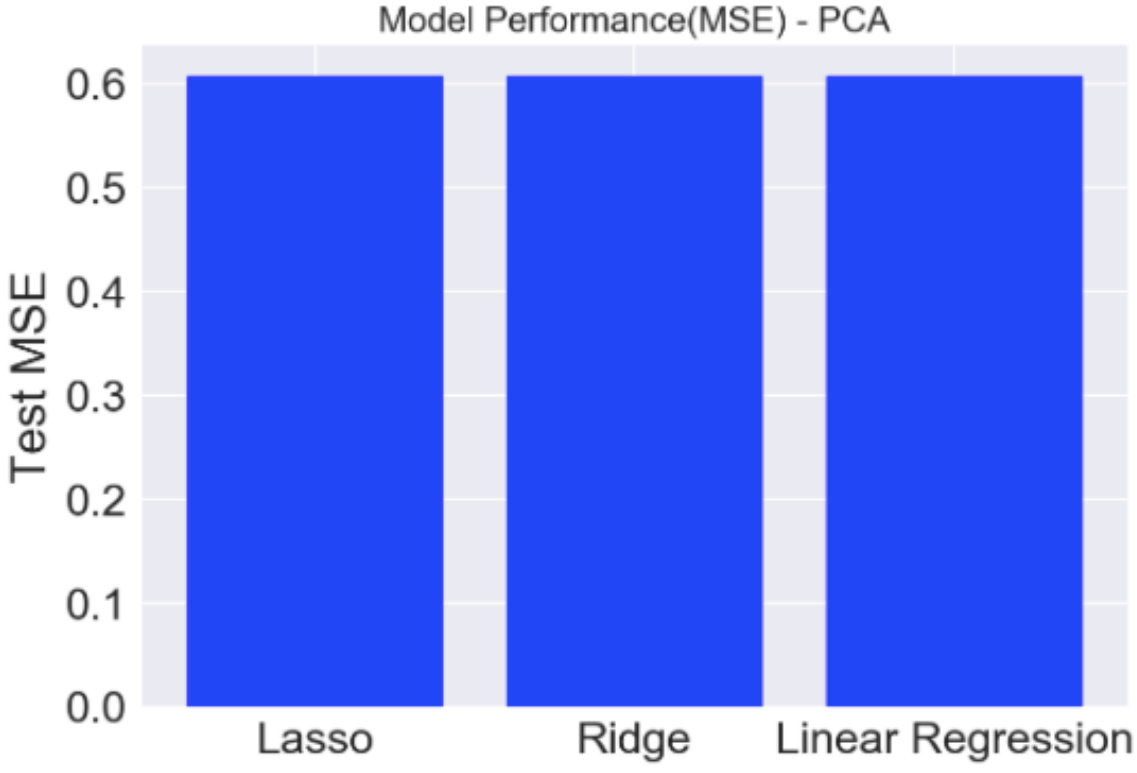
```
Hierarchical clustering:
Silhouette score: 0.204787168484
Homogeneity score: 0.181526740564
Completness: 0.696952074899
```

Linear Regression

All Variables



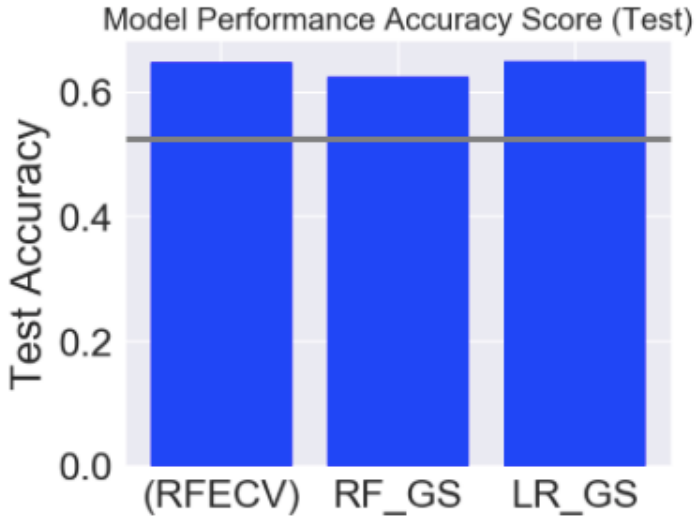
PCA



Classification

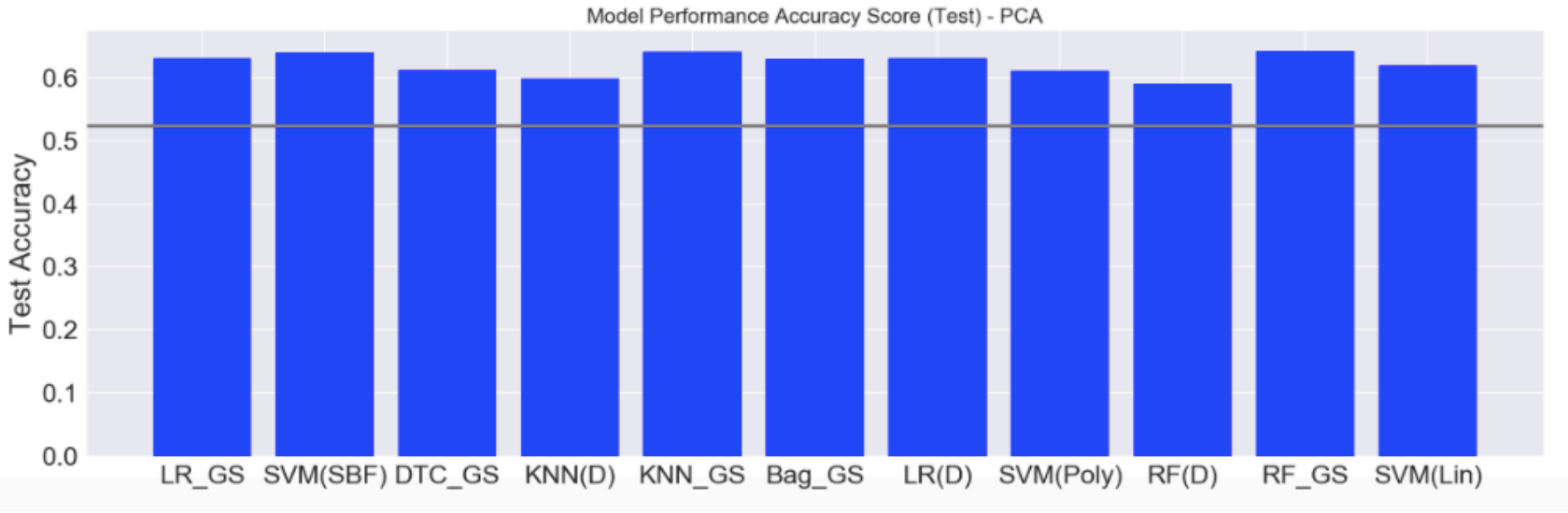
All Variables

```
Baseline Accuracy =  0.524
```



PCA

```
Baseline Accuracy =  0.524
```

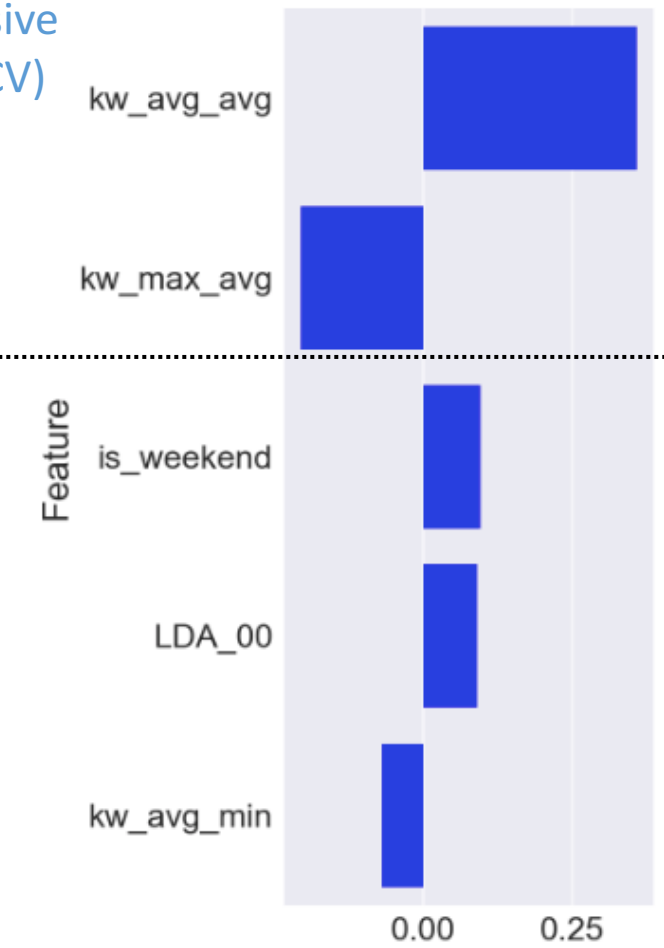


TOP 5 predictors related to size of news and article content

Most importance/Predictive Features

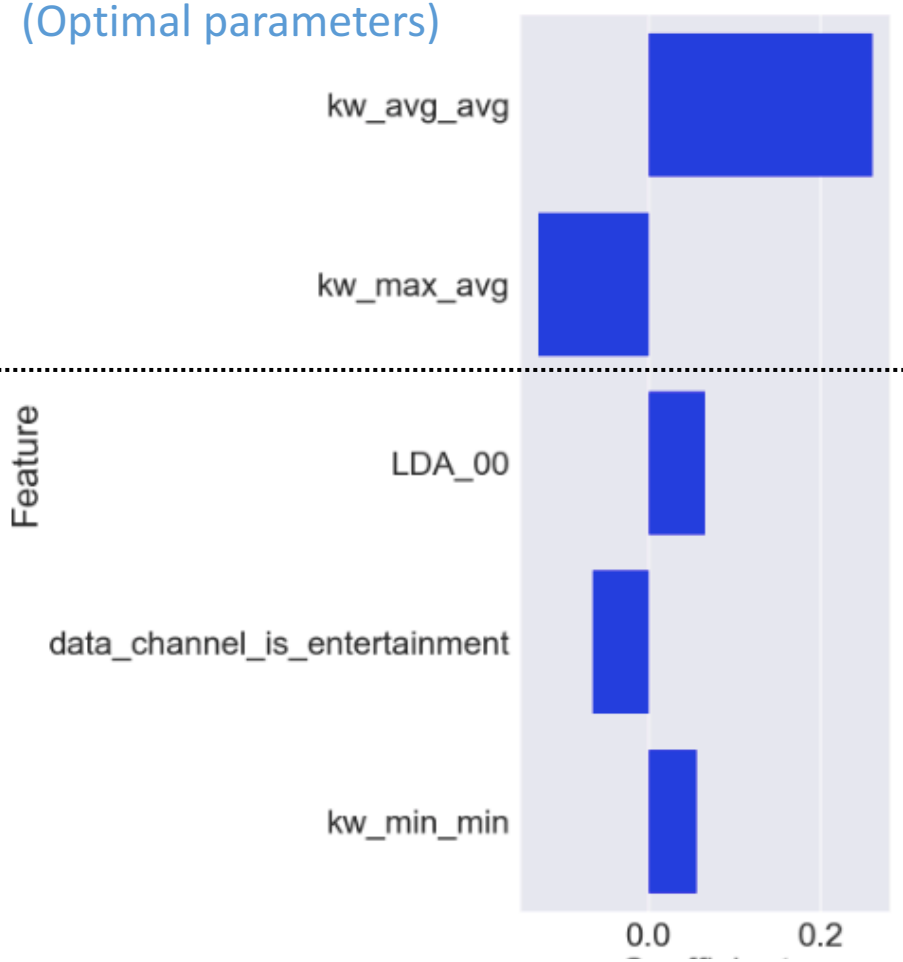
All Variables

Logistic Regression using recursive elimination (RFECV)



PCA

Random Forest (Optimal parameters)



NLP: Predicting with the article content

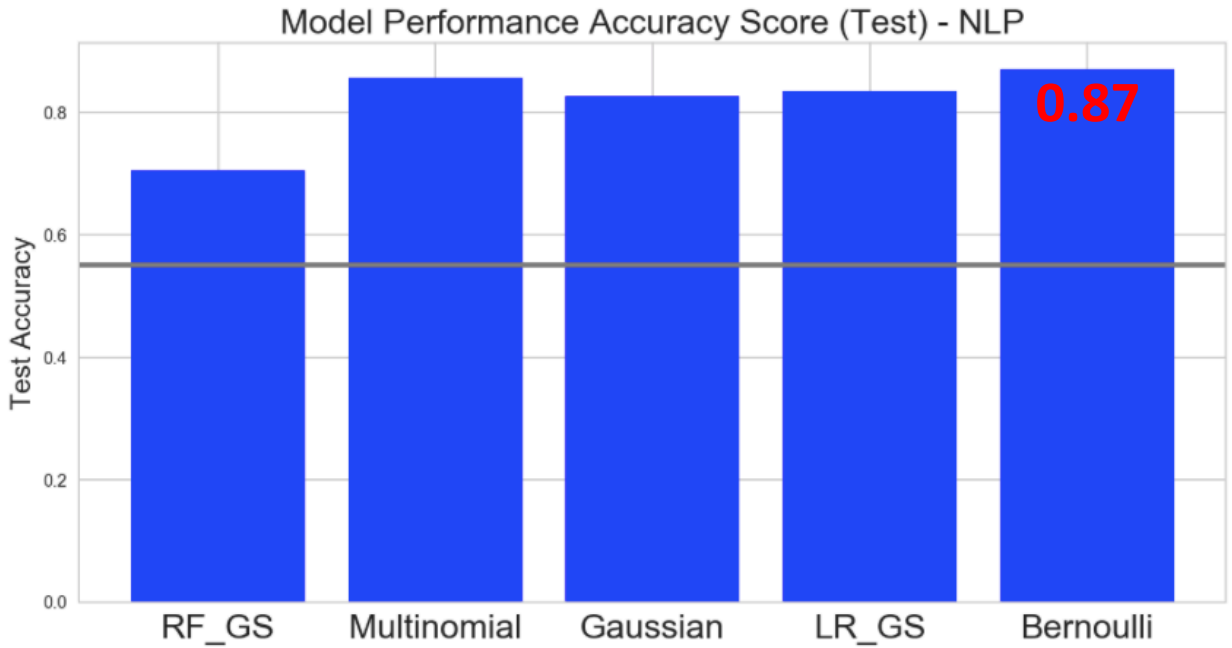
- Web Scraping / URL s columns
- Features extracted from the TITLE (TF-IDF), plus filtering using Select K best
- Features extracted from the BODY (TF-IDF), plus filtering using Select K best
- Modelling : Naive Bayes classifier
- Sentimental Analysis in order to find any insights relative to the content that appear in the title

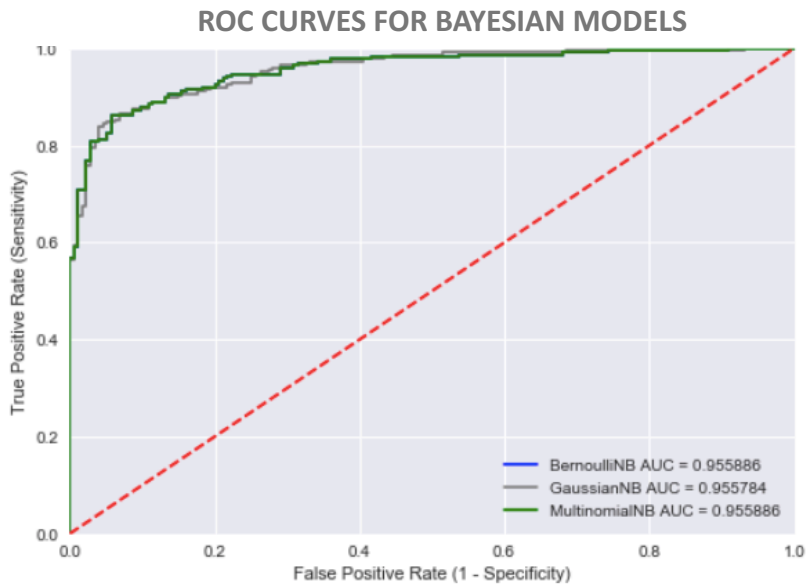
Based on 2000 articles of the news

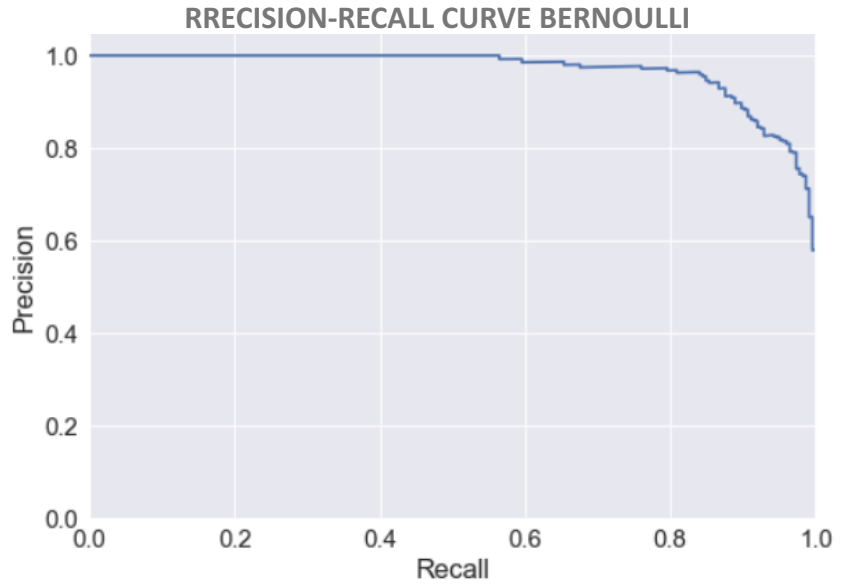
Accuracy score almost 90% - Bernoulli Model

NLP: Predicting with the article content

```
Baseline Accuracy =  0.551
```





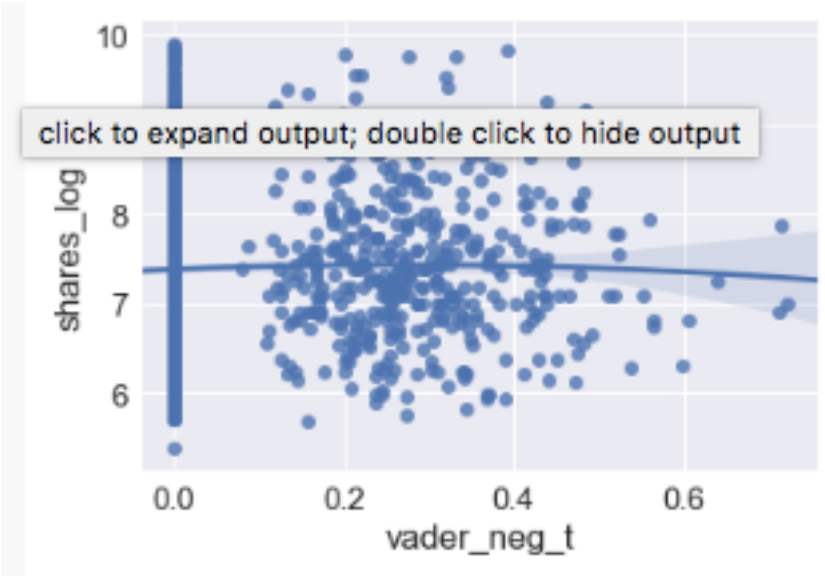


Let us what is happening with the title

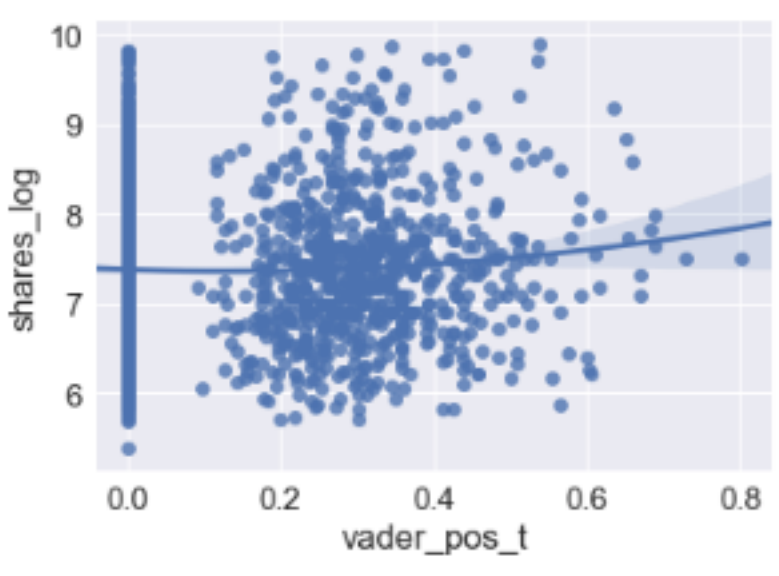
There is a positive correlation between social shares news and score positive in the title

NLP: sentimenal analisis

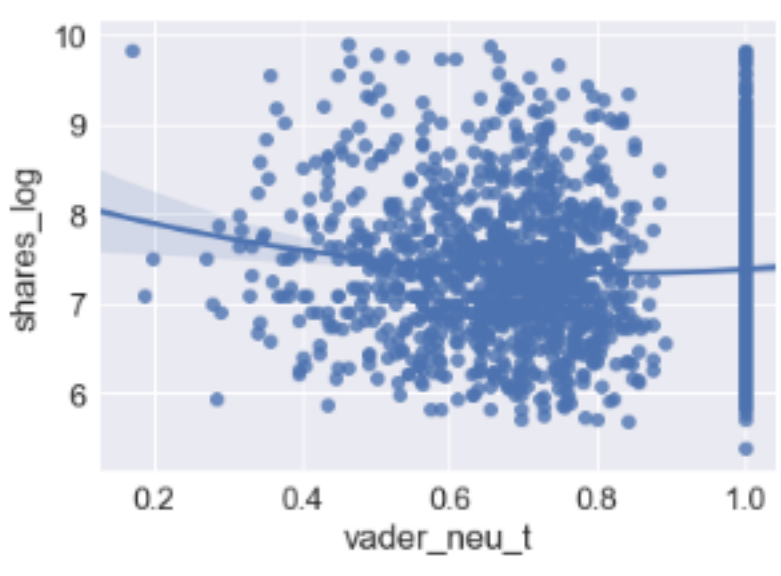
Score negative



Score positive



Score neutro



Based on 2000 articles of the news

Top 3 titles are including start with a digital numeral y two of them have video

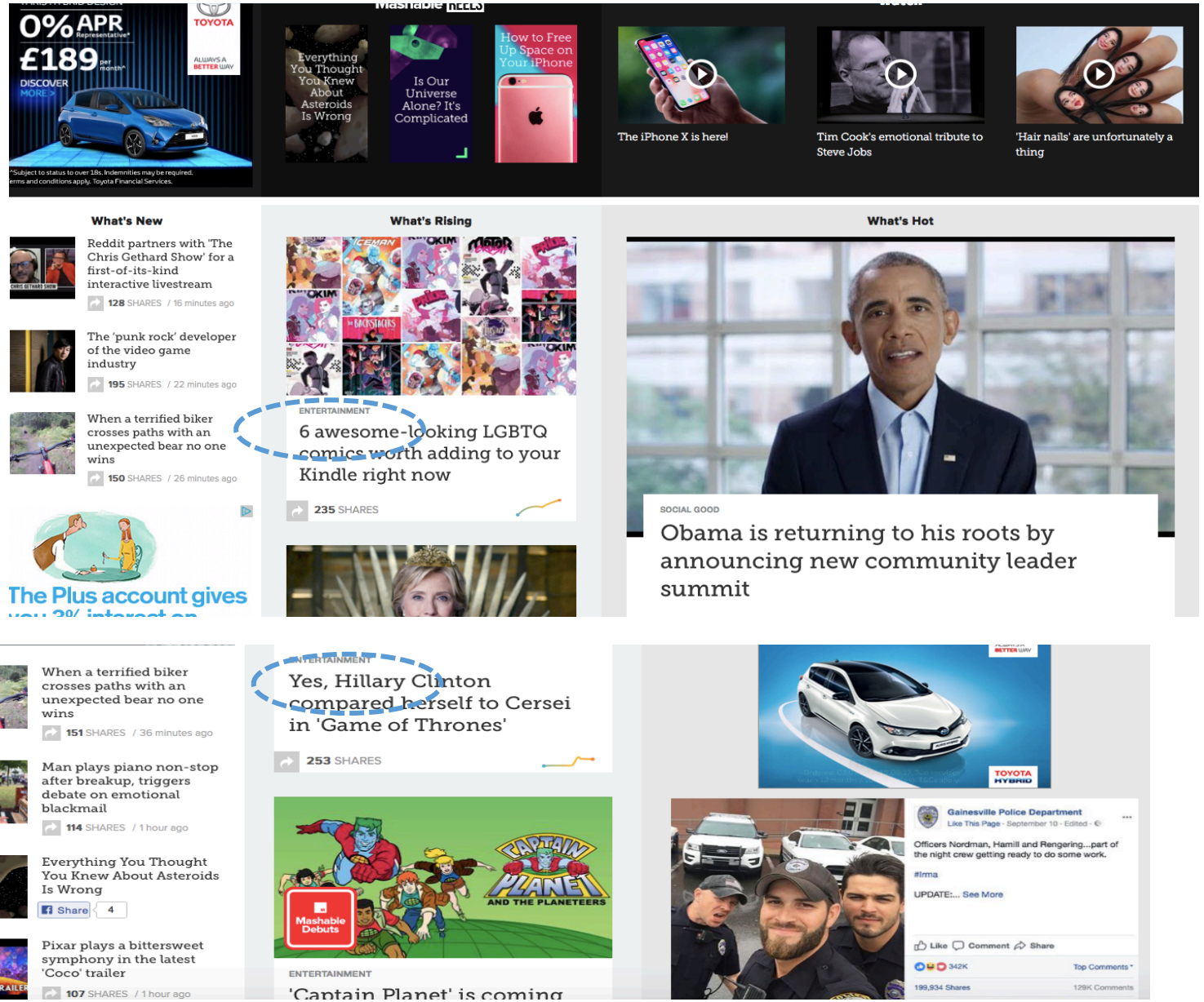
Positive words, strong visual and individual digit in title

Create THE BIGGEST IMPACT to share

Top 10 titles with positive score

'Love Is a Bracketfield' Facebook App Will Decide Best Romance Movie Ever
Puppies Adorably Predict Super Bowl Winner [VIDEO]
SAG Awards Recap: Best Moments and Acceptance Speeches
10 Awesome Pranks to Play On Your Facebook Friends
10 Best YouTube Channels for Free Fitness Videos
Government Wants to Create Free Public 'Super Wi-Fi'
The 10 Best Super Bowl Ads of All Time
The Best Super Bowl Ads in 60 Seconds [VIDEO]
Amy Poehler to Star in Best Buy Super Bowl Ad
Happy Superb Owl Sunday!

Based on 2000 articles of the news



Title:

- Strong visual
- Start with a digital numeral
- Start with positive imperative

Specific keywords may be more informative and tech

What words were Hot in titles

Conclusion: Predicting the popularity of online news

Recommendations

- Consideration the amount of the key words
- TITLE : Positive words, strong visual and digital numeral
- Articles of World and Entertainment
- Referenced articles in Mashable

Methodology

- The models provided here demonstrate that utilizing a two-class label system with a partition at the median provides the best models and including analysis the article content.

- For future works, I would like to carry out in-depth NLP and Text Analysis.