

Mashable

Media and entertainment company for super fans

CHANNELS

Video

Entertainment

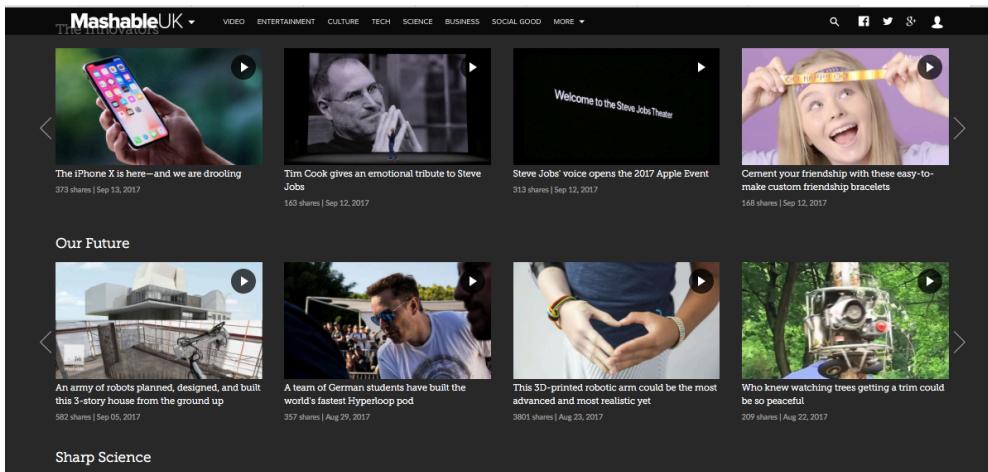
Culture

Tech

Science

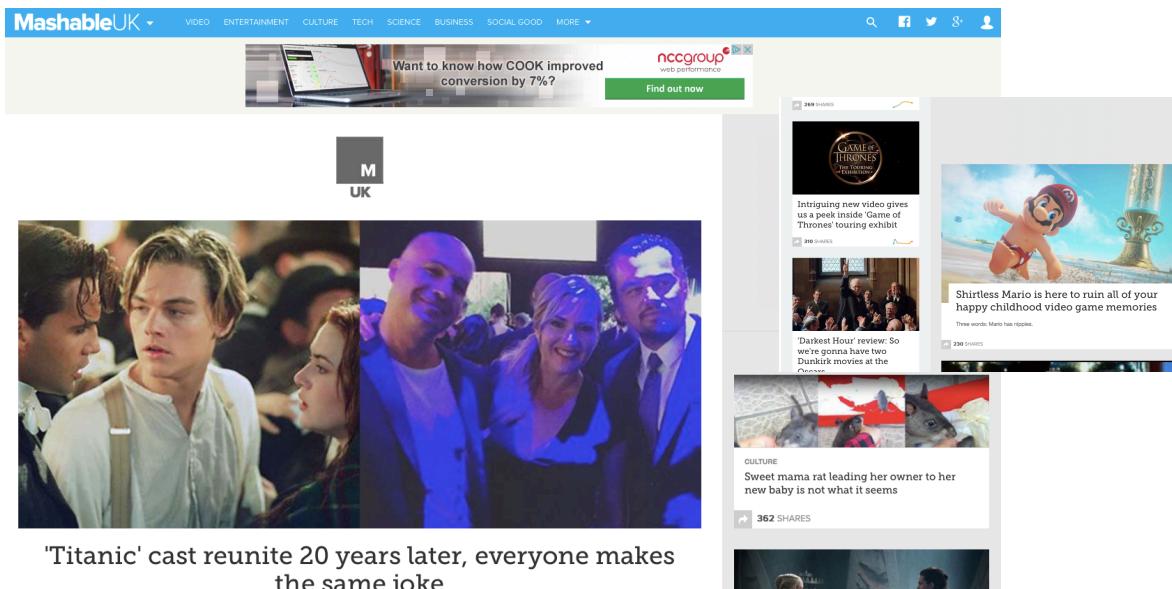
*We know **the future of TV** looks nothing like the past. Great TV won't be made for mass audiences. It'll be made for the right audiences, **using data** both to inspire creativity and **connect shows with influential viewers**," said Pete Cashmore, Founder and CEO of Mashable. "It won't happen on 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



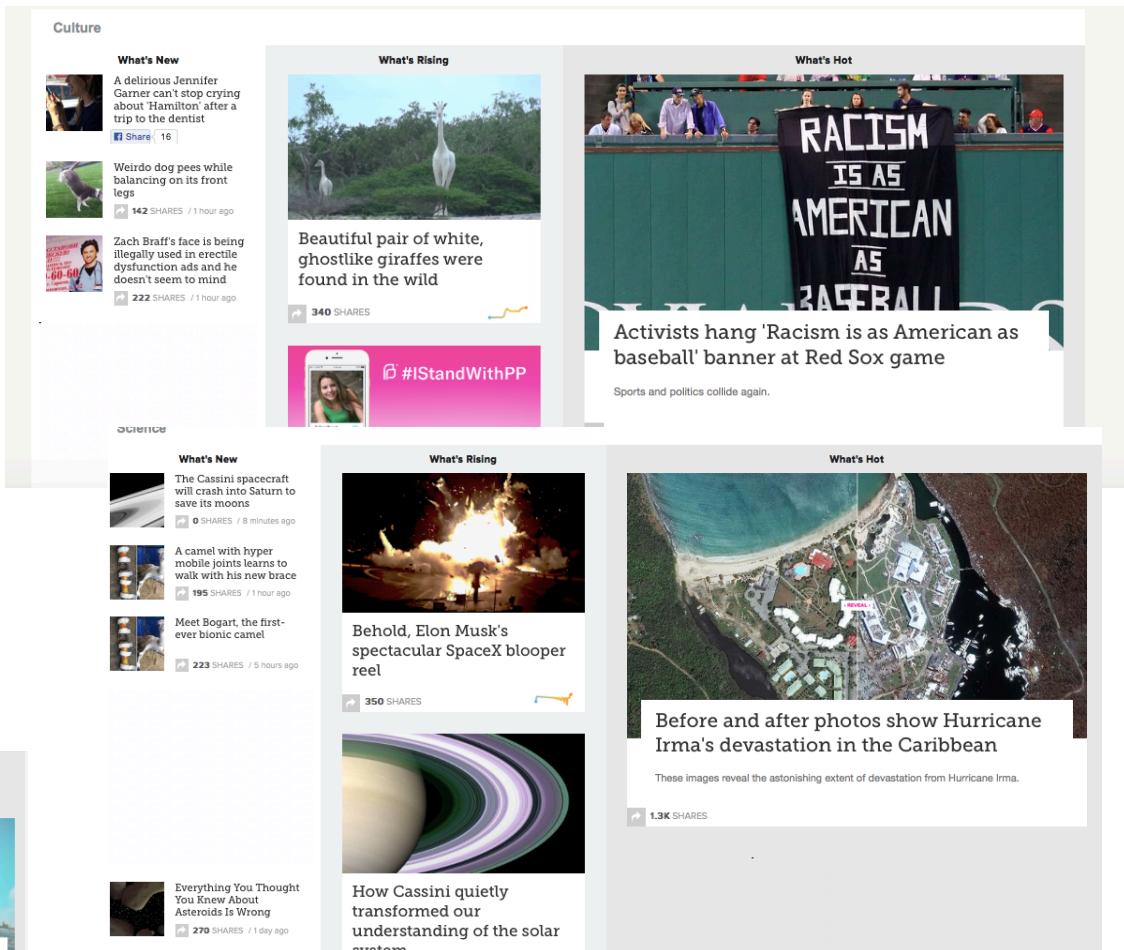
The MashableUK video section features a grid of video thumbnails. The top row includes: "The iPhone X is here—and we are drooling" (373 shares), "Tim Cook gives an emotional tribute to Steve Jobs" (168 shares), "Steve Jobs' voice opens the 2017 Apple Event" (313 shares), and "Cement your friendship with these easy-to-make custom friendship bracelets" (148 shares). Below this is a section titled "Our Future" with four videos: "A army of robots planned, designed, and built this 3-story house from the ground up" (582 shares), "A team of German students have built the world's fastest Hyperloop pod" (357 shares), "This 3D-printed robotic arm could be the most advanced and most realistic yet" (3801 shares), and "Who knew watching trees getting a trim could be so peaceful" (209 shares).

Entertainment



The MashableUK entertainment section includes a banner for NCC Group web performance. Below it, a large image shows Leonardo DiCaprio and Kate Winslet from the movie "Titanic". A caption below reads: "'Titanic' cast reunite 20 years later, everyone makes the same joke". To the right, there are several video thumbnails: "Intriguing new video gives us a peek inside 'Game of Thrones' touring exhibit" (269 shares), "Darkest Hour' review: So we're gonna have two Dunkirk movies at the Oscar" (210 shares), "Shirtless Mario is here to ruin all of your happy childhood video game memories" (230 shares), and "Sweet mama rat leading her owner to her new baby is not what it seems" (362 shares).

Culture



The MashableUK culture section is divided into several sections: "Culture", "What's New", "What's Rising", "Science", and "What's Hot". The "Culture" section has a "What's New" article about Jennifer Garner (16 shares) and a "What's Rising" article about white giraffes (340 shares). The "Science" section has a "What's New" article about the Cassini spacecraft (0 shares) and a "What's Rising" article about a bionic camel (350 shares). The "What's Hot" section features a large image of activists holding a "RACISM IS AS AMERICAN AS BASEBALL" banner at a Red Sox game, with a caption: "Activists hang 'Racism is as American as baseball' banner at Red Sox game". Another "What's Hot" article shows satellite images of Hurricane Irma's impact (1.3K shares).

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

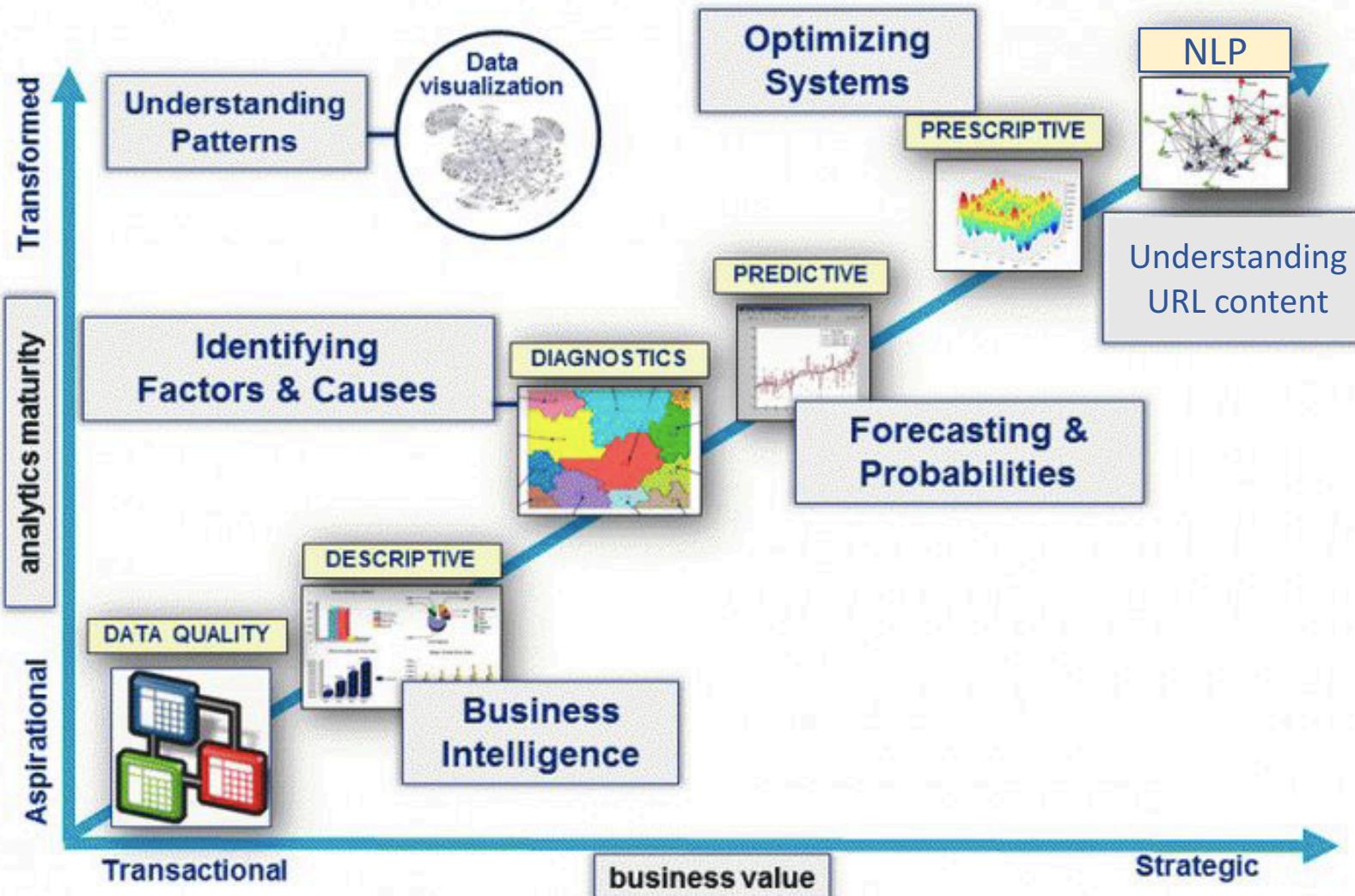
<https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

- Number of Attributes 58
- Number of Records 39664

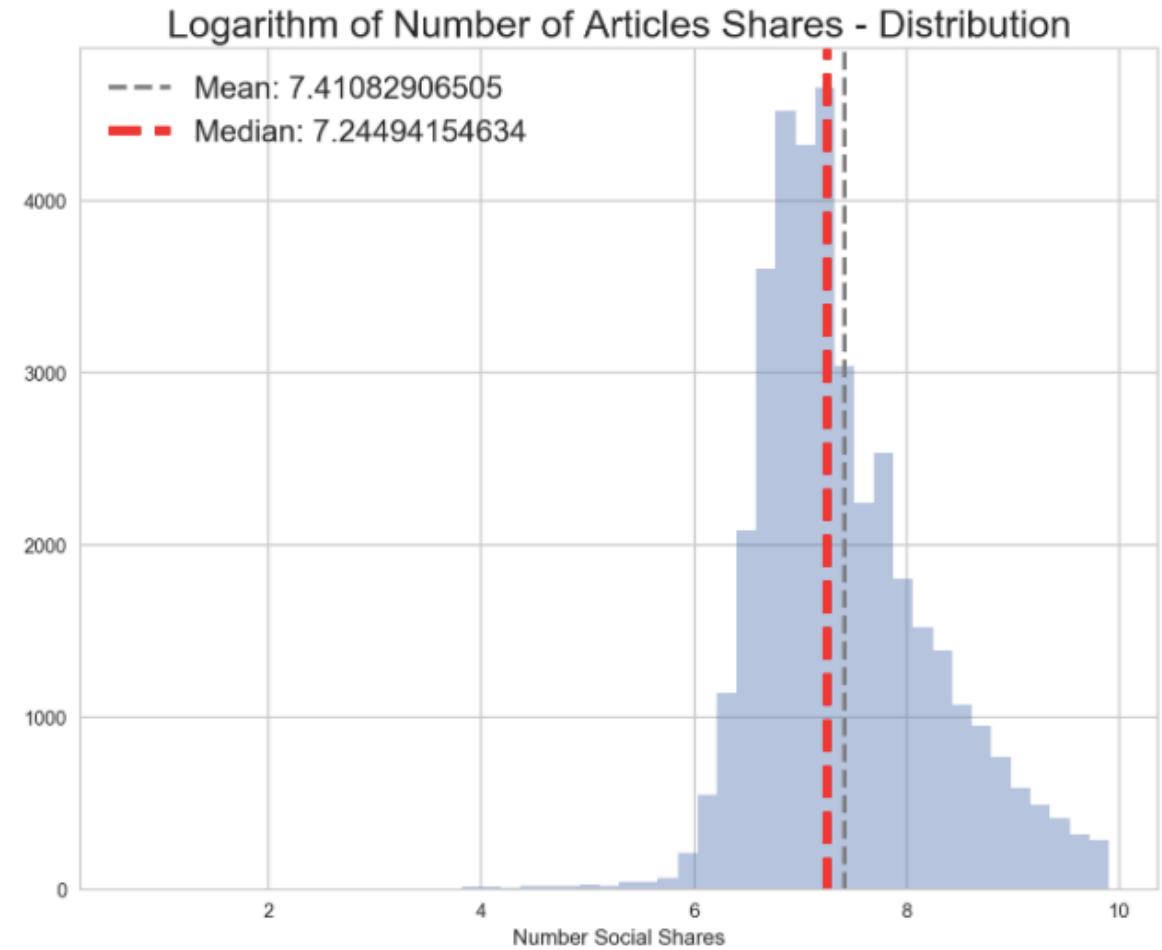
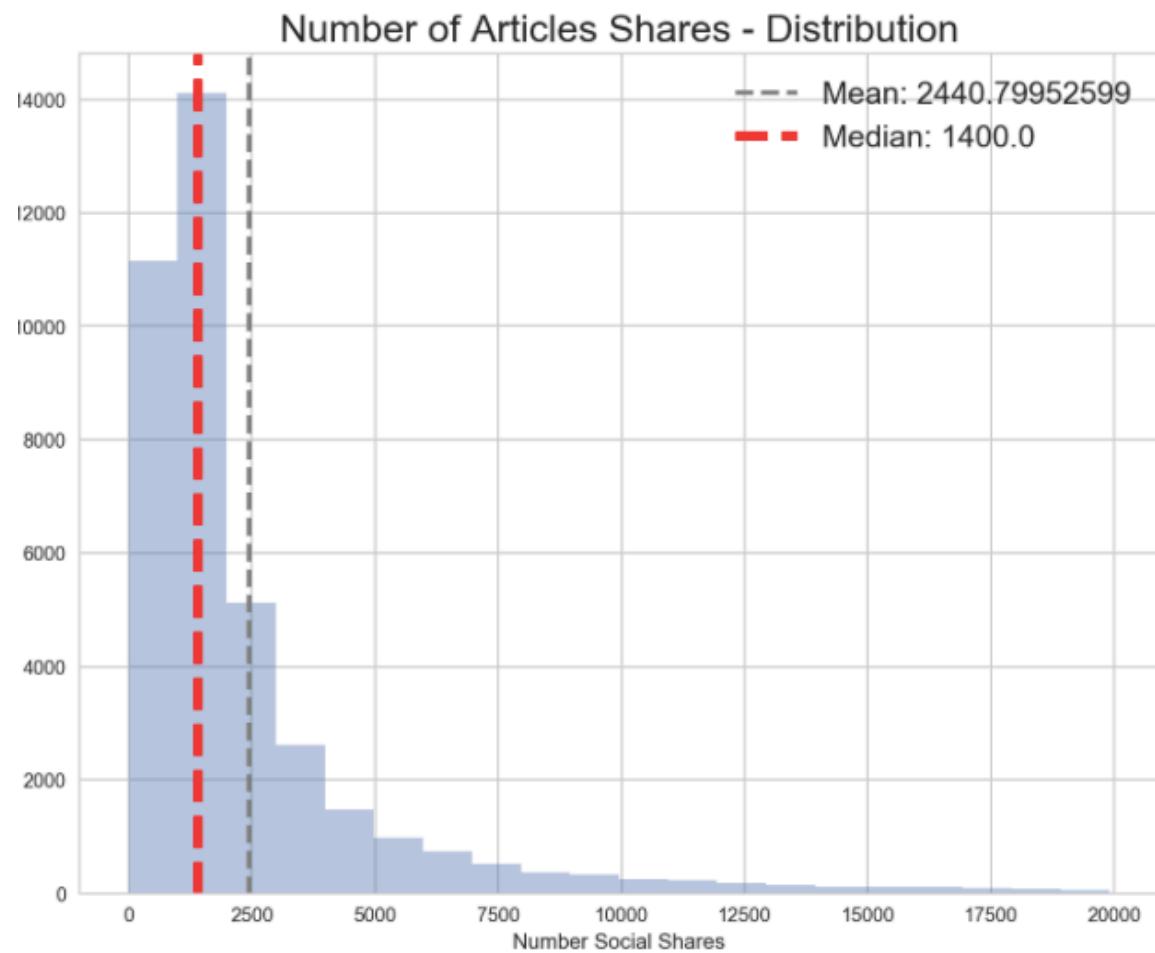
b) Web Scraping: Importing a dataset from the web and saving as a local PostgreSQL

c) Importing a dataset from the URL of the article

Data Analysis

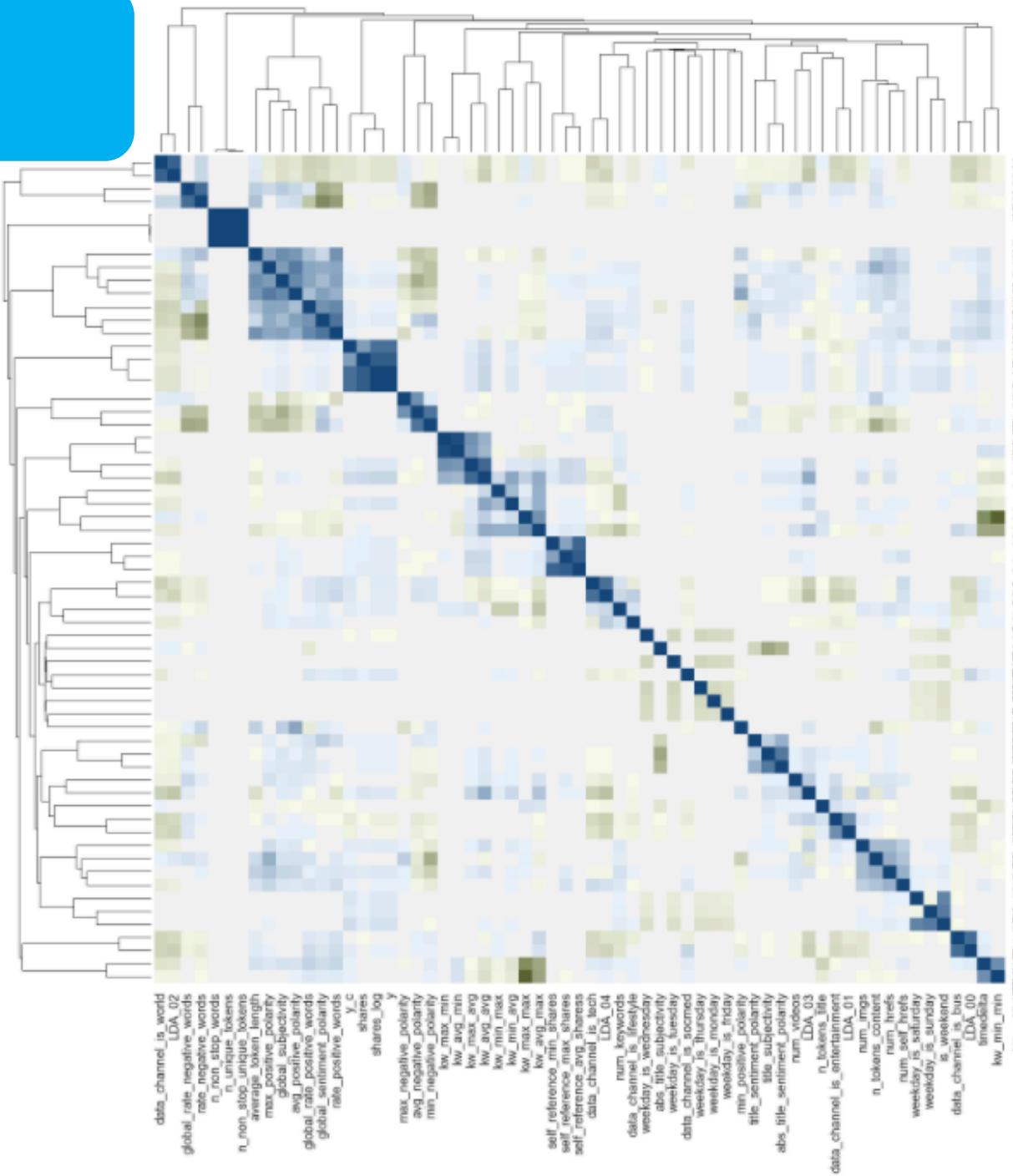


Target

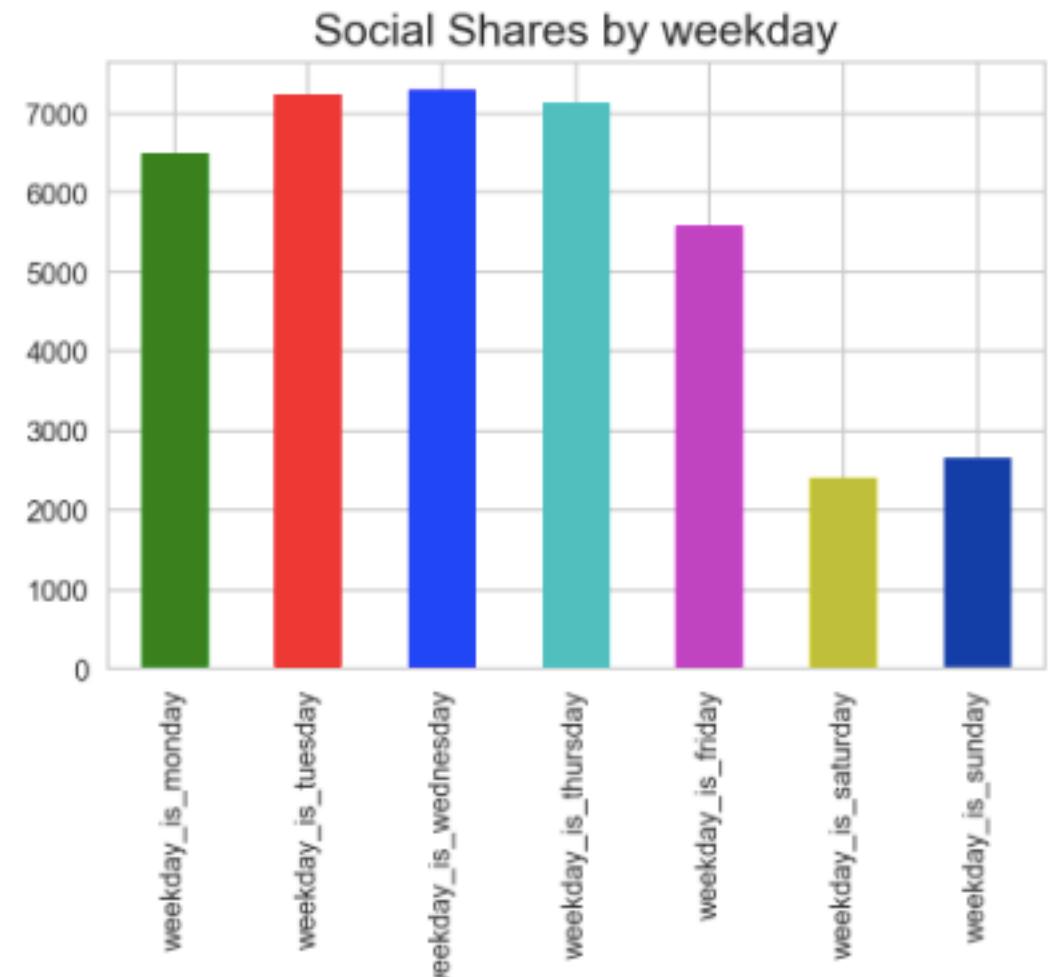
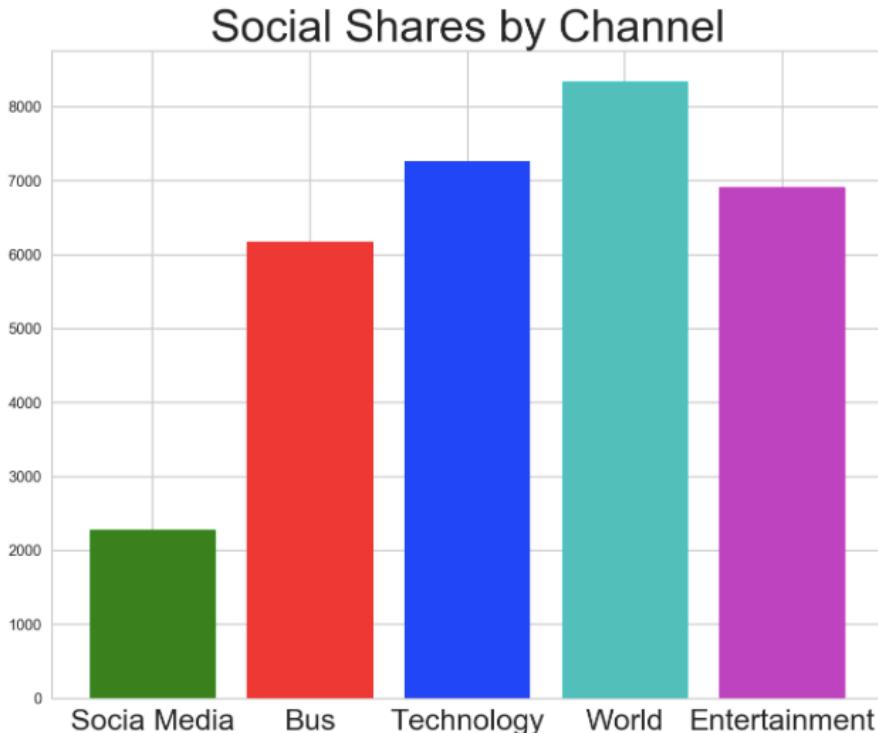


Correlation

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_shares 6287.758556 5998.758401
```

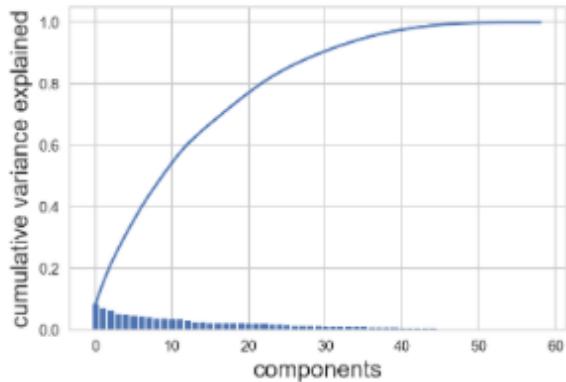
Caveats and limitations in the data set

- 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

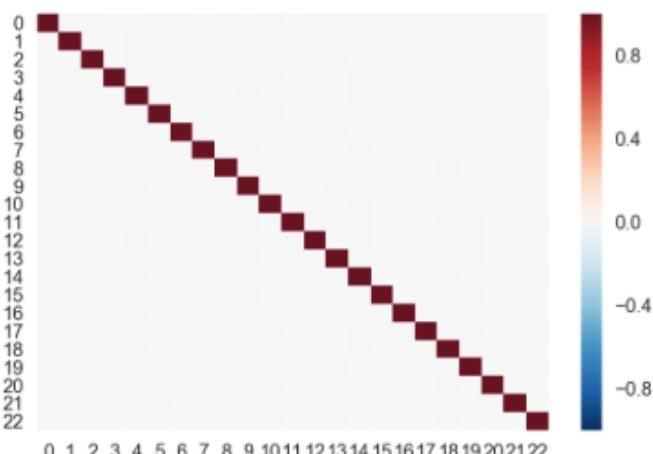
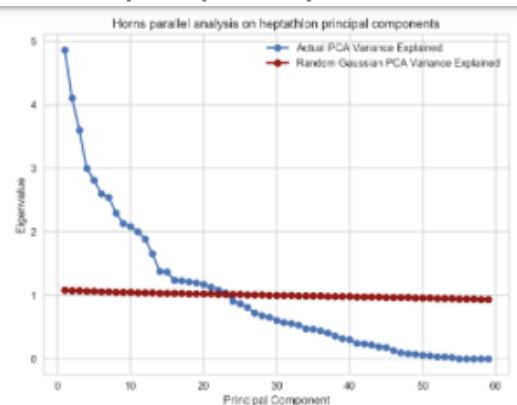
Almost 80% of the explained variance can be explained by just 23 of the principal components

Dimensionality reduction

cumulative variance explained vs components



Horns parallel analysis on heptathlon principal components



Clustering Analysis

Number Cluster = 3

K-Means

KMeans clustering:

Silhouette score: 0.207201788072

Homogeneity score: 0.195510046802

Completness: 0.735793987952

Hierachical clustering

Hierarchical clustering:

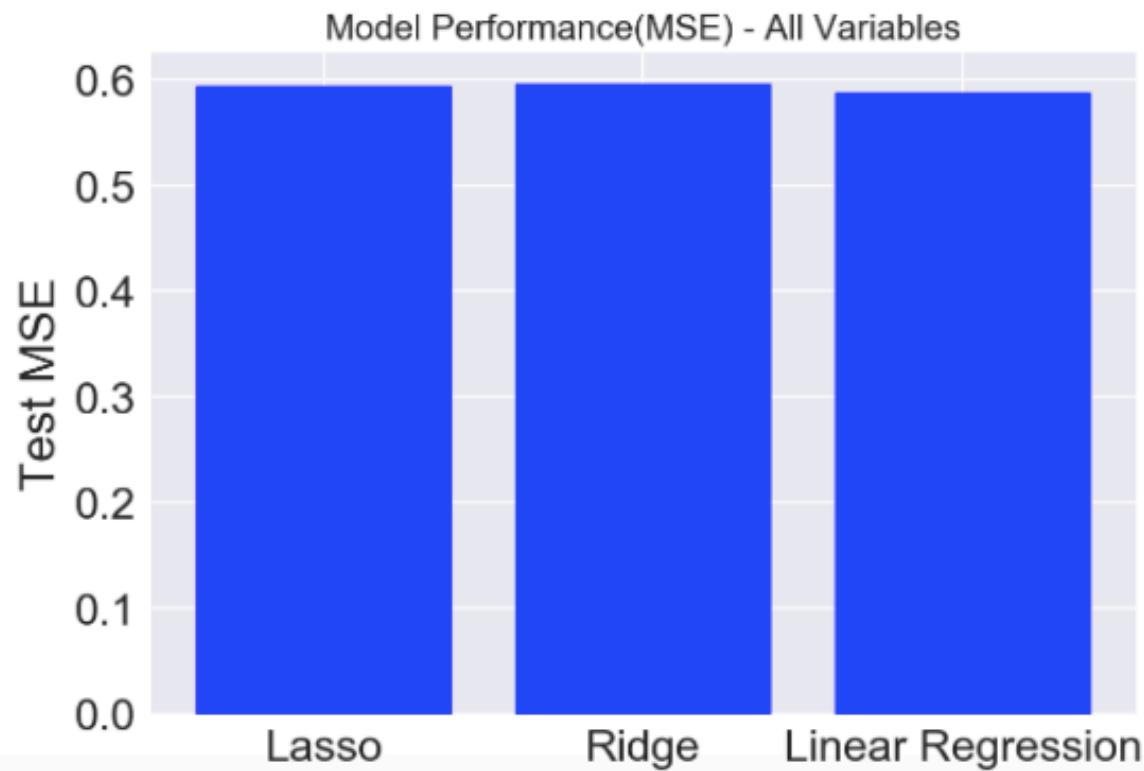
Silhouette score: 0.204787168484

Homogeneity score: 0.181526740564

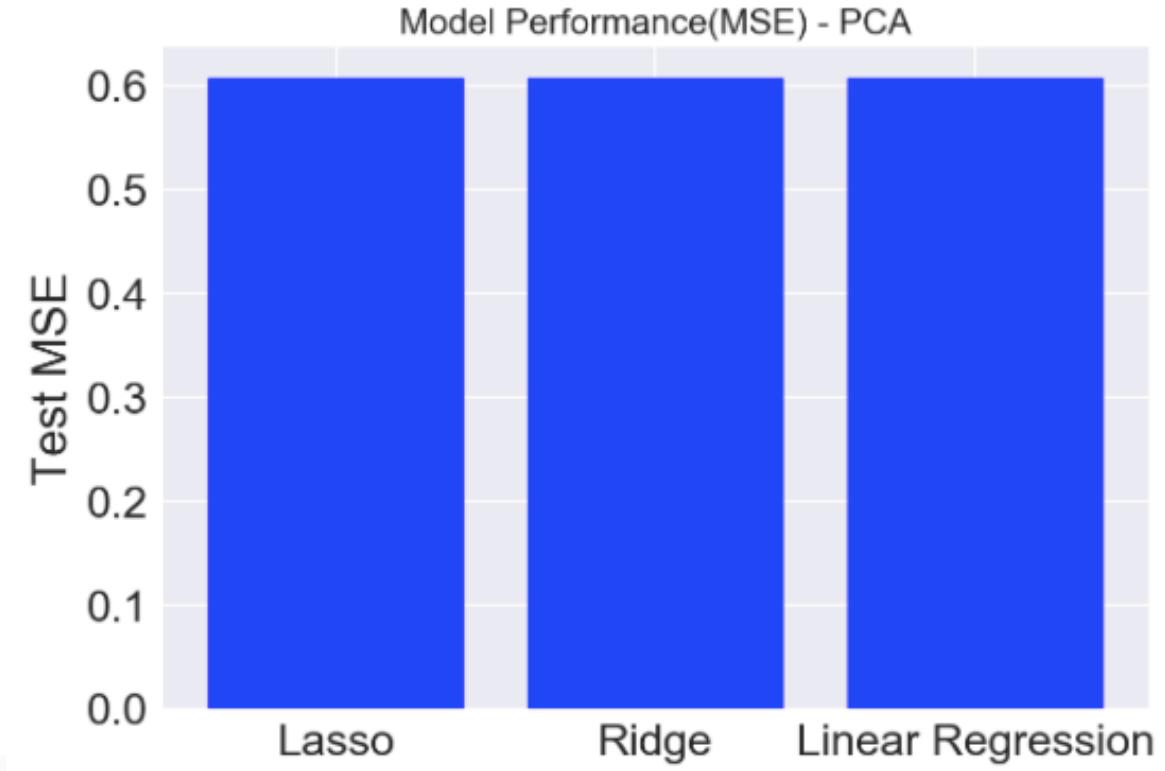
Completness: 0.696952074899

Linear Regression

All Variables



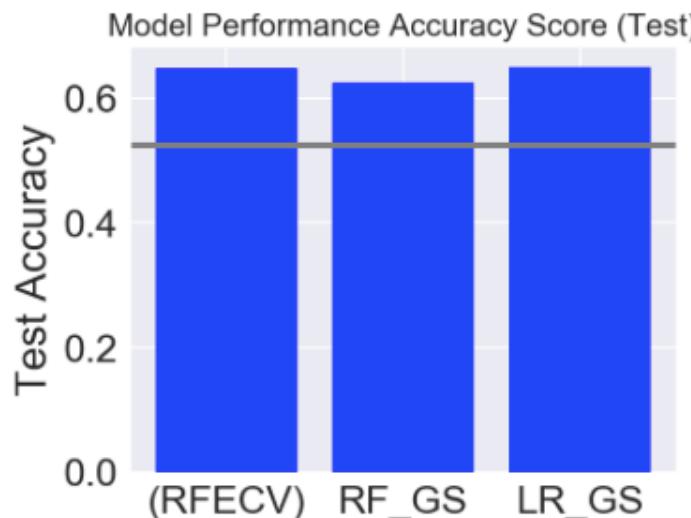
PCA



Classification

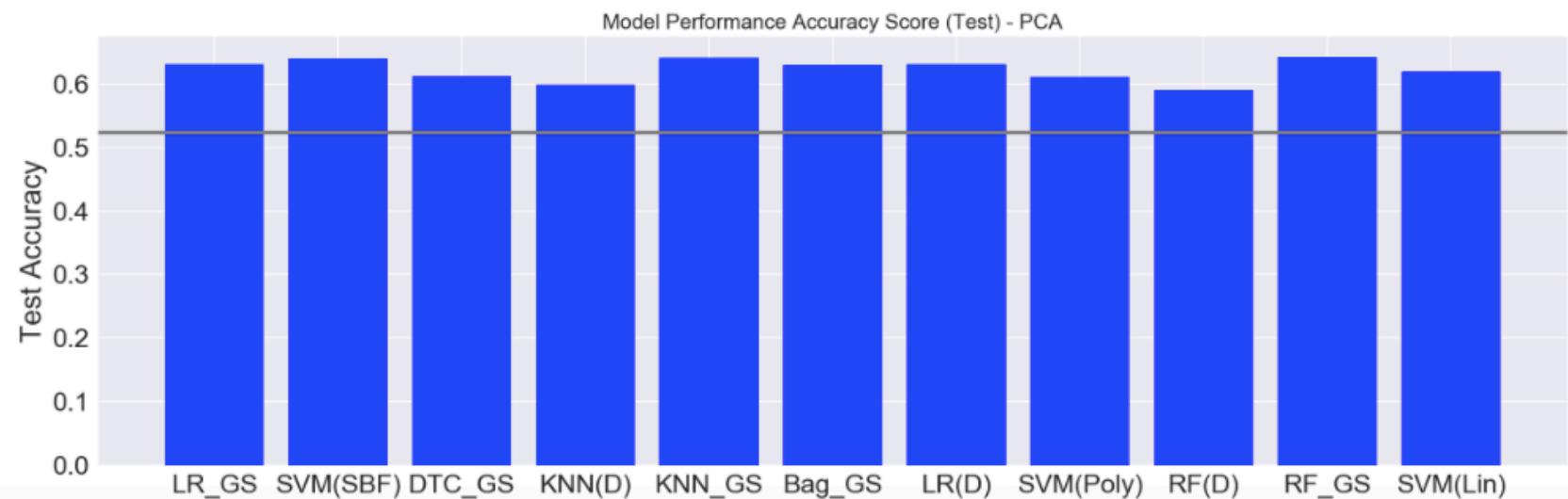
All Variables

Baseline Accuracy = 0.524



PCA

Baseline Accuracy = 0.524

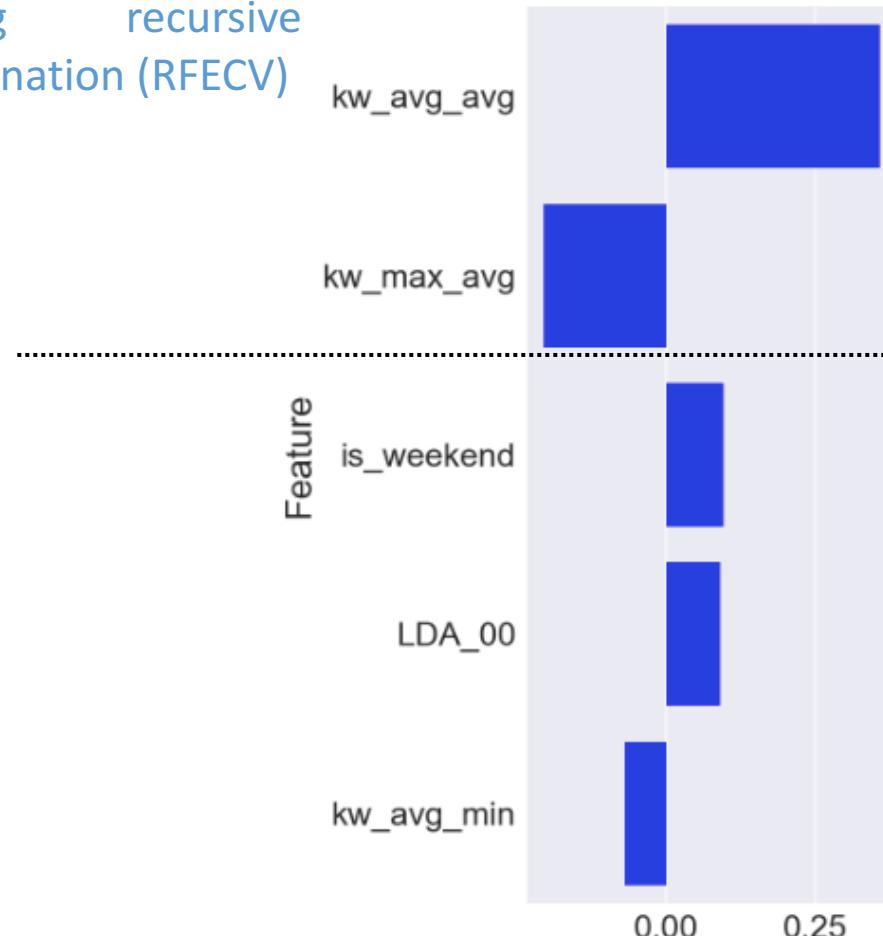


TOP 5 predictors related to size of news and article content

Most importance/Predictive Features

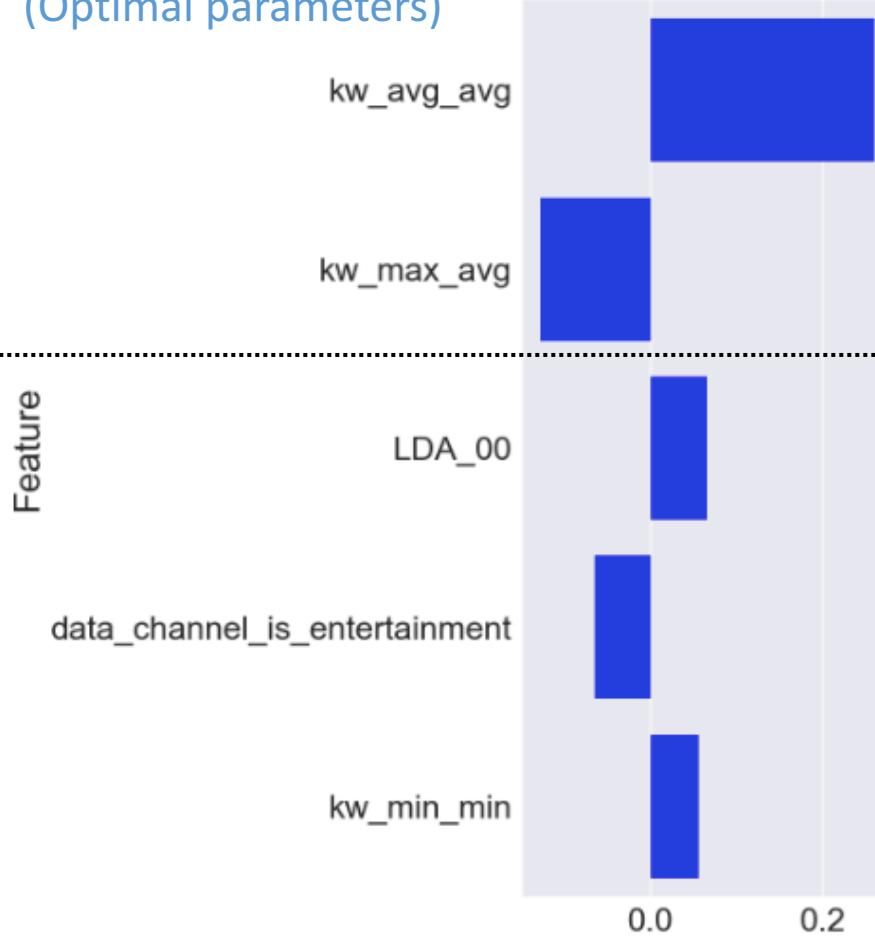
Logistic Regression using recursive elimination (RFECV)

All Variables



Random Forest (Optimal parameters)

PCA



NLP: Predicting with the article content

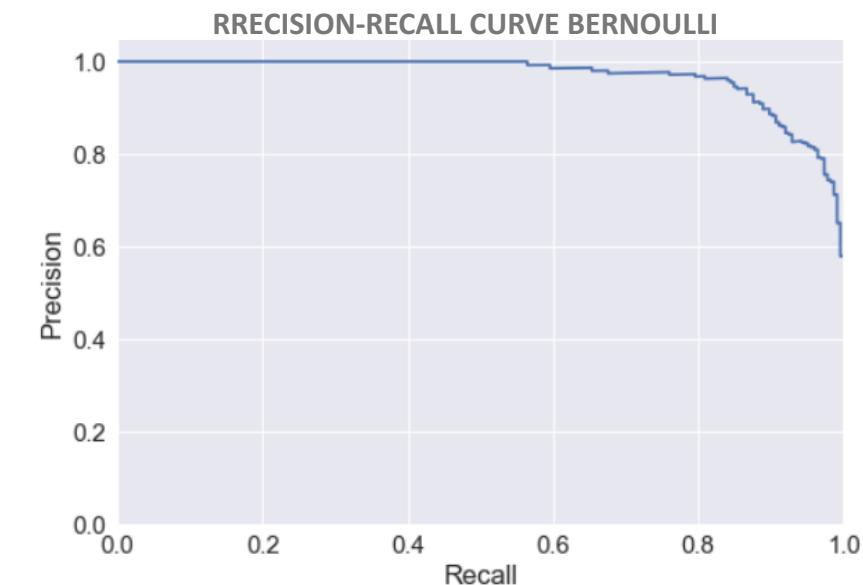
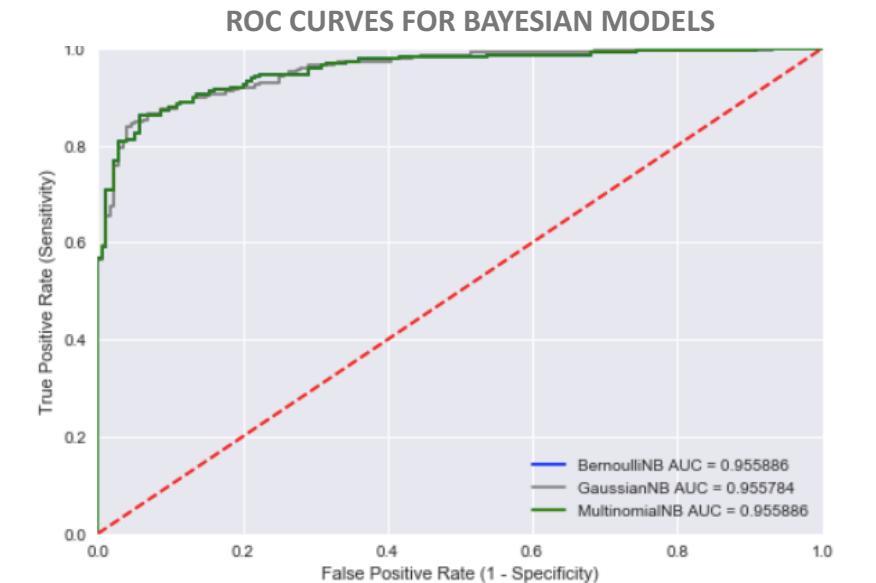
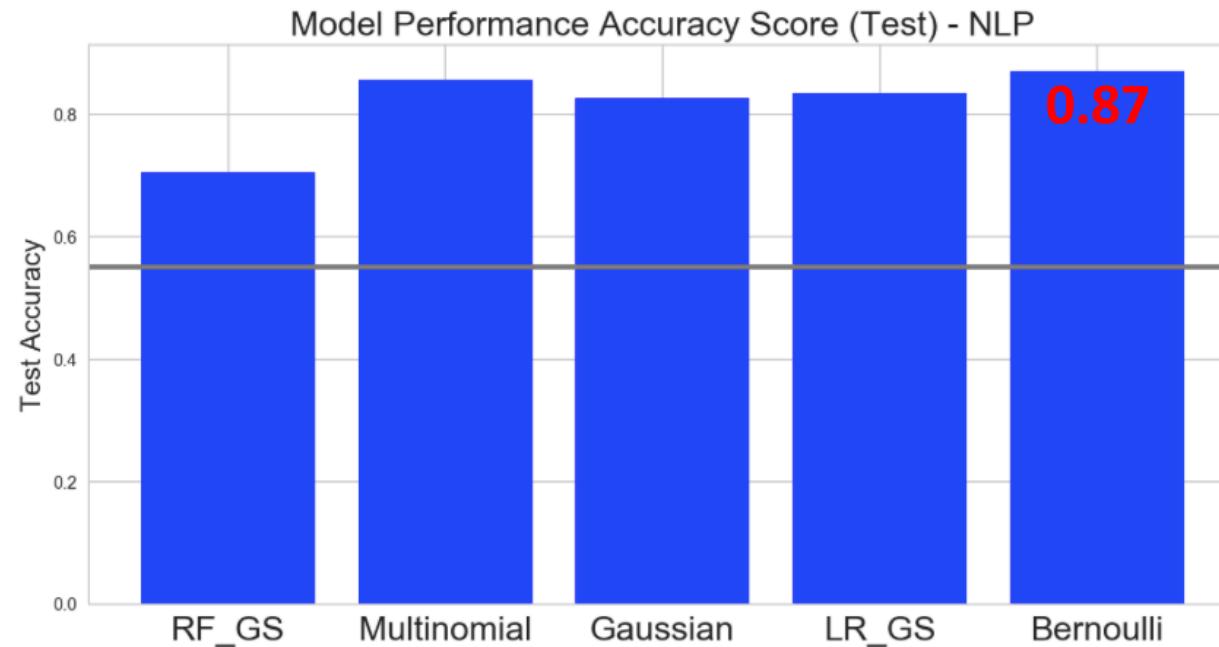
- ❖ Web Scraping / URL s columns
- ❖ Features extracted from the TITLE (TF-IDF), plus filtering using Select K best
- ❖ Features extracted from the BODY (TF-IDF), plus filtering using Select K best
- ❖ Modelling : Naive Bayes classifier
- ❖ Sentimental Analysis in order to find any insights relative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



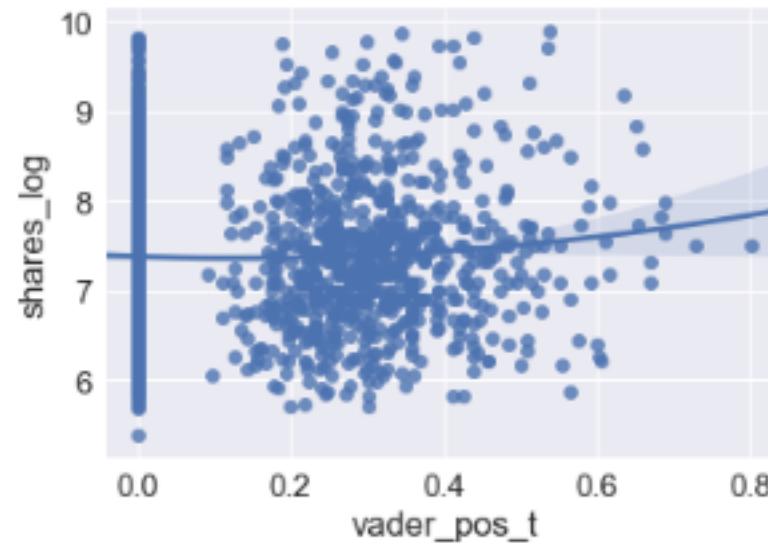
There is a positive correlation between social shares news and score positive in the title

NLP: sentinel analisis

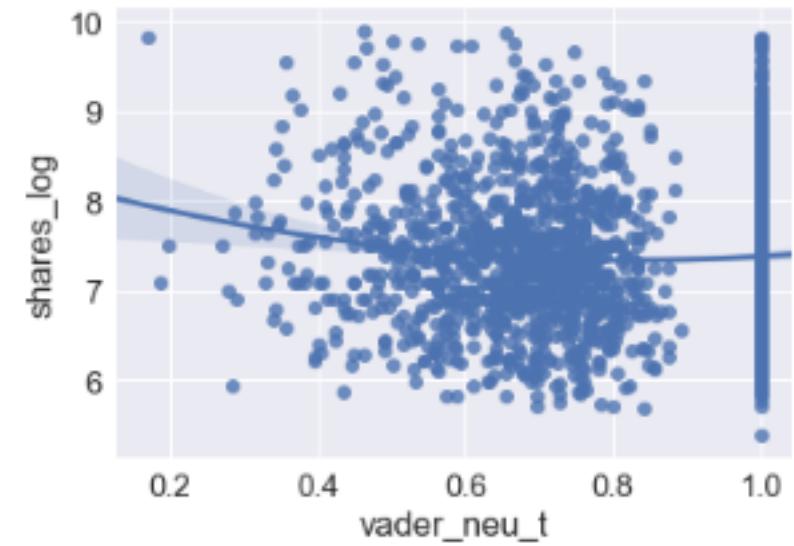
Score negative



Score positive



Score neutro



Based on 2000 articles of the news

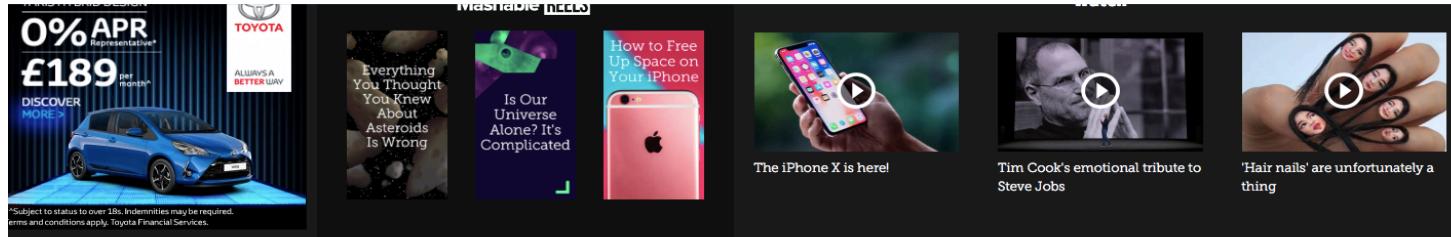
Top 3 titles are including start with a digital numeral y two of them have video

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!



Title:

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

What's New

Reddit partners with 'The Chris Gethard Show' for a first-of-its-kind interactive livestream
128 SHARES / 16 minutes ago

The 'punk rock' developer of the video game industry
195 SHARES / 22 minutes ago

When a terrified biker crosses paths with an unexpected bear no one wins
150 SHARES / 26 minutes ago

The Plus account gives you 20% interest on...

When a terrified biker crosses paths with an unexpected bear no one wins
151 SHARES / 36 minutes ago

Man plays piano non-stop after breakup, triggers debate on emotional blackmail
114 SHARES / 1 hour ago

Everything You Thought You Knew About Asteroids Is Wrong
[Share](#) 4

Pixar plays a bittersweet symphony in the latest 'Coco' trailer
107 SHARES / 1 hour ago

What's Rising

ENTERTAINMENT 6 awesome-looking LGBTQ comics worth adding to your Kindle right now
235 SHARES

ENTERTAINMENT Yes, Hillary Clinton compared herself to Cersei in 'Game of Thrones'
253 SHARES

ENTERTAINMENT 'Captain Planet' is coming

What's Hot

SOCIAL GOOD Obama is returning to his roots by announcing new community leader summit

TOYOTA HYBRID Toyota's new hybrid car is the best way to get around town

Gainesville Police Department Officers Nordman, Hamill and Rengering...part of the night crew getting ready to do some work.

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.