greatlearning

ONLINE NEWS POPULARITY

Group 3:
Abhisar Narkhede
Ameer Khan
Deepak Shende
Sneharth Bhajani
Sharmistha Ghosh
Vishnu .P.S

Under the guidance of:
Mr. Srikar Muppidi

HOW MARKETING STRATEGY GOT CHANGED

Before **Now**

How do Blogging websites generate revenue?

Introduction

- The dataset summarizes a set of features about articles published by Mashable, a well-known news website over a period of two years from Jan 2013 - Jan 2015
- The objective is to predict the number of shares depending on the features if the article to be published would be popular on the internet or no.
- 39,644 observations
- 61 attributes
- No missing values, but some topics were unclassified
- Target: number of shares

- The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content be publicly accessed and retrieved using the provided urls.
- Acquisition date: January 8, 2015

Mashable

What is Mashable!

MashableUK VIDEO ENTERTAINMENT CULTURE TECH SCIENCE BUSINESS MORE

- 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."

What is the goal?

PREDICTING THE POPULARITY OF ONLINE NEWS

Based on number of social shares of articles

Data description and Problem Statement

Data Source :

- Mashable

Problem Statement:

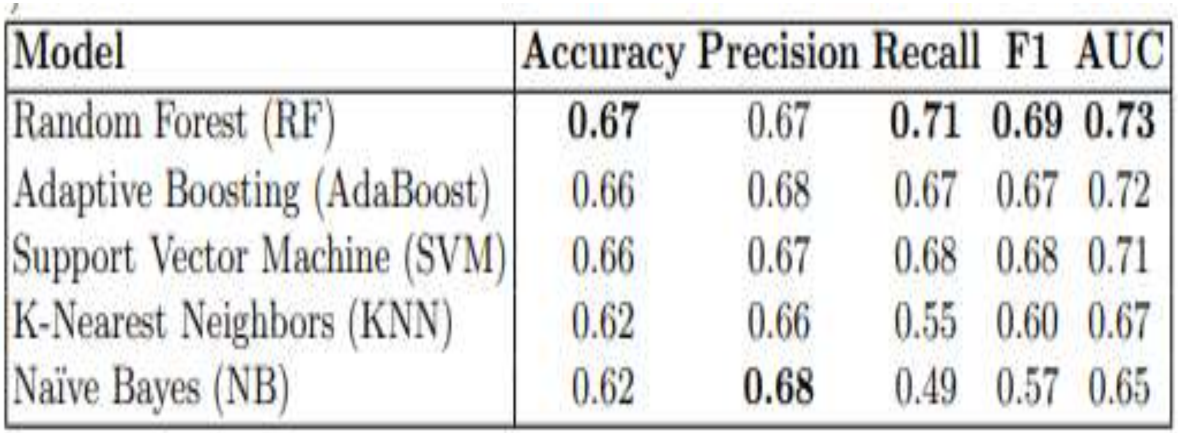
Problem :

- Even though the content of article is good still few articles don't get good number of shares and displaying Ads with it doesn't make profit.

Solution :

- Popularity prediction of online article aims to predict the future popularity of new article prior to the publication estimating the number of shares, likes and comments that particular article will get depending on various features.
- Depending on that we can display Ads with respect to the popularity.

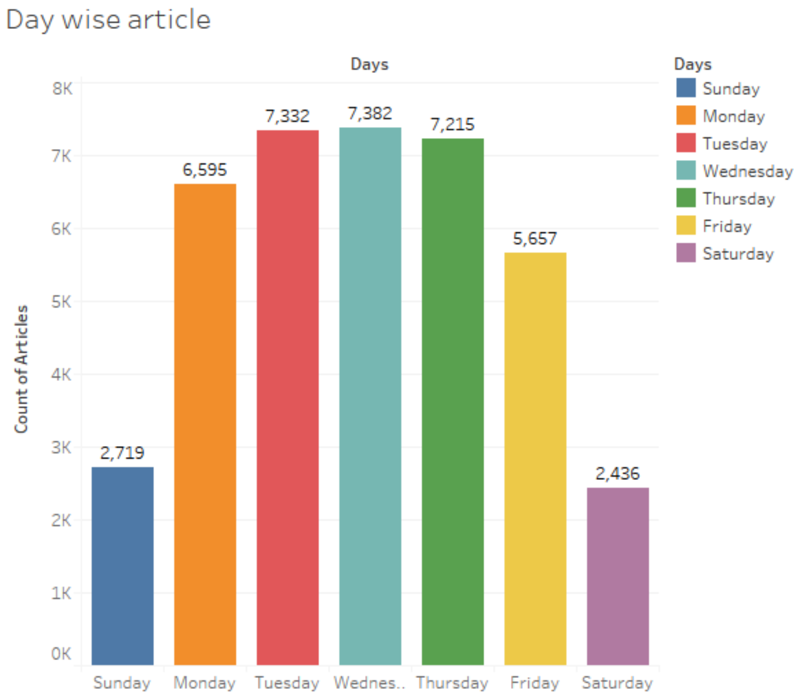
Please give the benefit of your findings and reasons for your conclusion!!!!

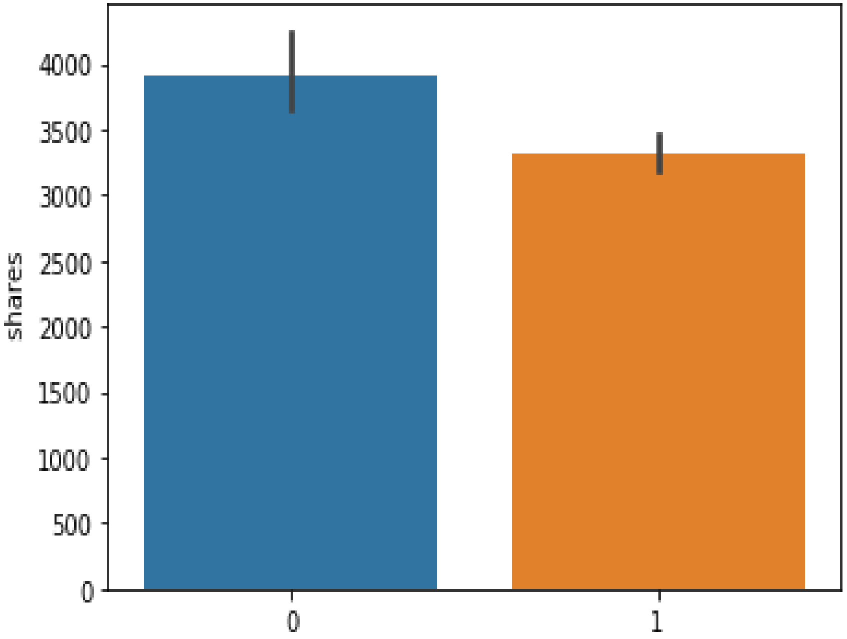


Model	Accuracy	Precision	Recall	F1	AUC
Random Forest (RF)	**0.67**	0.67	**0.71**	**0.69**	**0.73**
Adaptive Boosting (AdaBoost)	0.66	0.68	0.67	0.67	0.72
Support Vector Machine (SVM)	0.66	0.67	0.68	0.68	0.71
K-Nearest Neighbors (KNN)	0.62	0.66	0.55	0.60	0.67
Naïve Bayes (NB)	0.62	**0.68**	0.49	0.57	0.65

- Research Paper: ‘**A Proactive IDSS for Predicting the Popularity of Online News’**
- Objective : To obtain better accuracy
- Link: http://repositorium.sdum.uminho.pt/bitstream/1822/39169/1/main.pdf

Exploratory Data Analysis



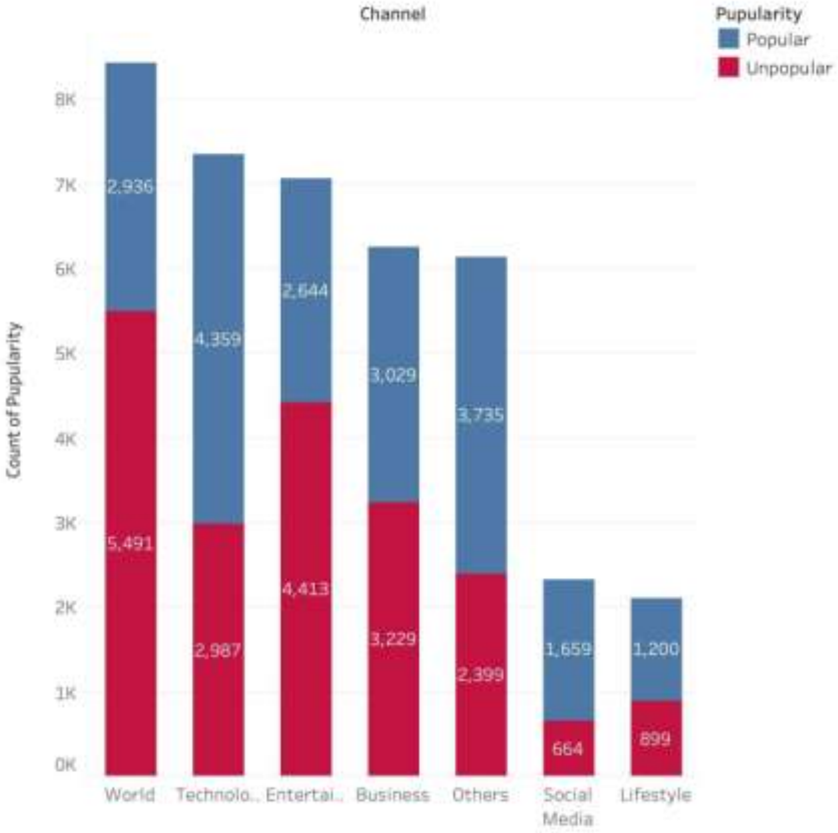


- **Wednesday** has maximum articles getting released
- Weekdays have maximum number of shares

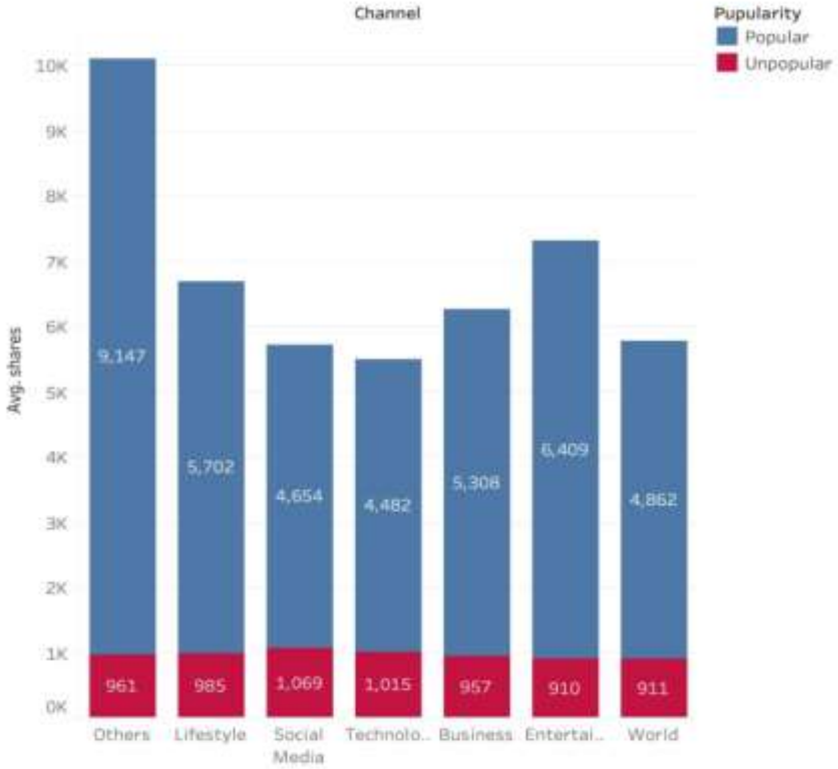
- Weekday (0) : More share
- Weekend (1): Less Share

DAY WISE SHARE

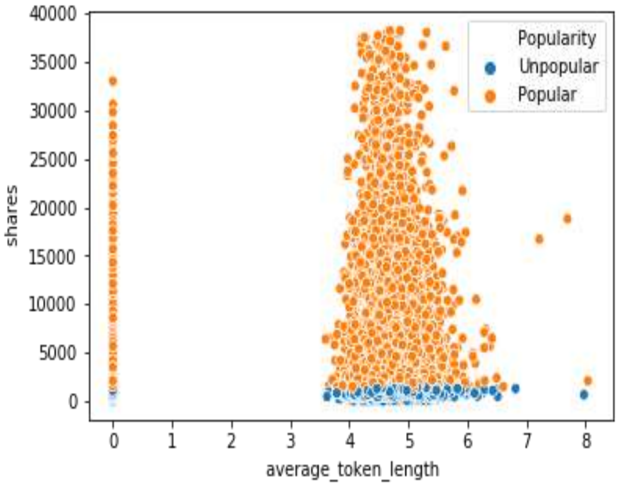
Popularity based on the type of Channel

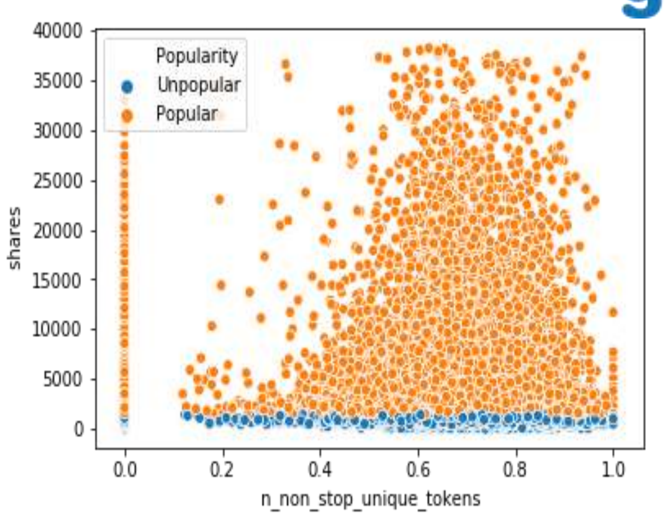


Average number of Shares of a Channel based on Popularity

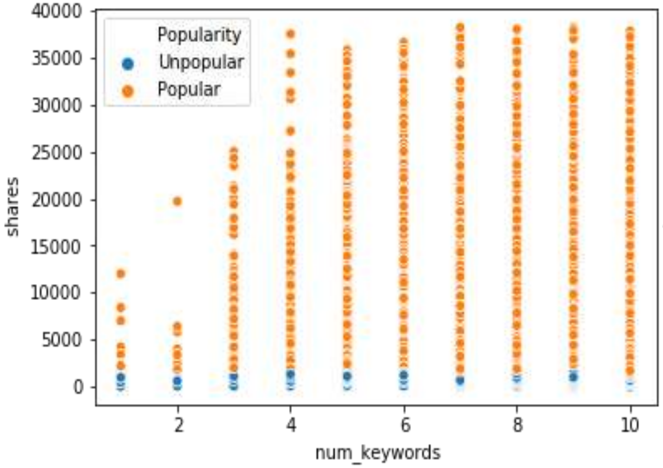


- More number of articles are getting published under data channel “**Technology**”.
- “**Others**” has more number of shares. Apart from that ‘**Entertainment**’ has more shares.

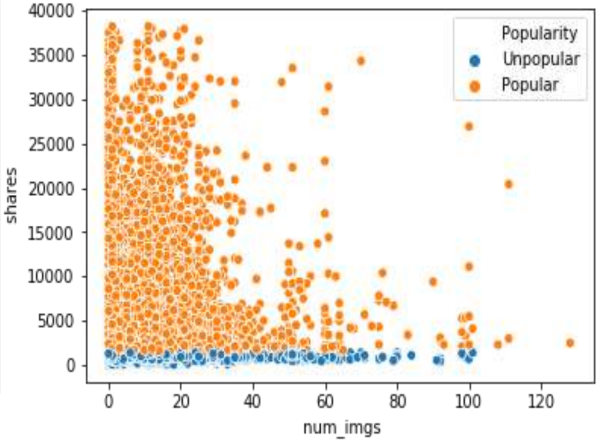




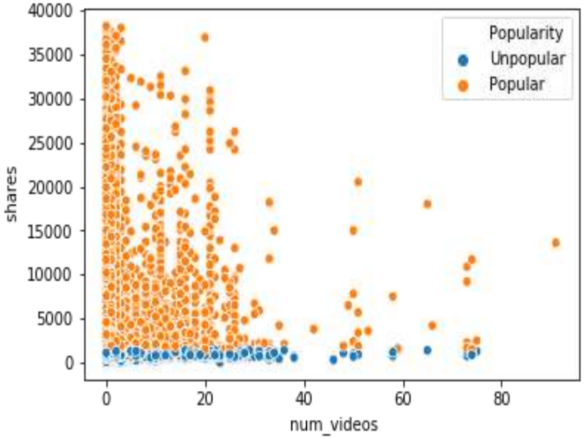
- Rate of frequency of unique words is between 0.4 – 0.8
- Rate of frequency of non-stop unique words is between 0.5 – 0.9
- Average length of the words in content is between 4 - 5.5



- More number of keywords articles are getting shared more
- In our case it ranges between 4 to 10



- Image count range between 1-25
- Lesser the image sharing is high

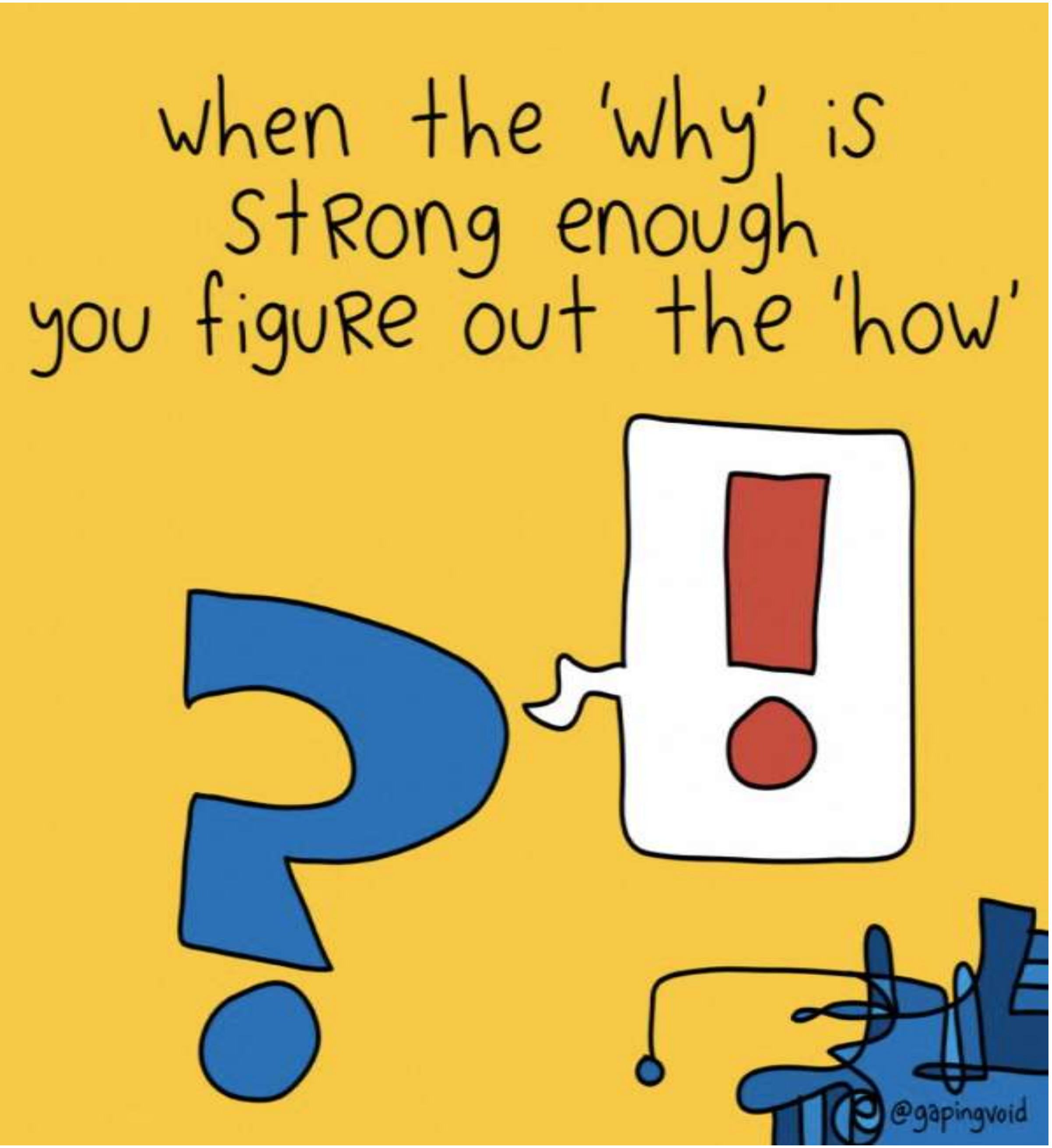


- Video count range between 1-4
- Lesser the videos sharing is high

Observations

- The **number of keywords** in the metadata influences the shares to a margin.The higher

  the value the better the shares chances. A value upward of 5 is recommended.
- The **content** should have less than 1500 words. The lesser the better.
- **Title** should be between 6 - 17 words.
- **Unique words** should be between 0.3 - 0.8%
- **No. of links** between 1 and 40 is preferred.
- **Images** - 0 to 3
- **Videos** - 0 to 25
- Minimal images and videos will make an article more interesting
- More articles are getting published on World data channel.
- **Lifestyle** and **entertainment** based articles are preferred more by people.
- Best popular articles are usually posted on **Monday** and **Wednesday** (and a bit of Tuesdays).
- **Weekends** are not preferred to publish an article.


when the 'why' is
strong enough
you figure out the 'how'
@gapingvoid

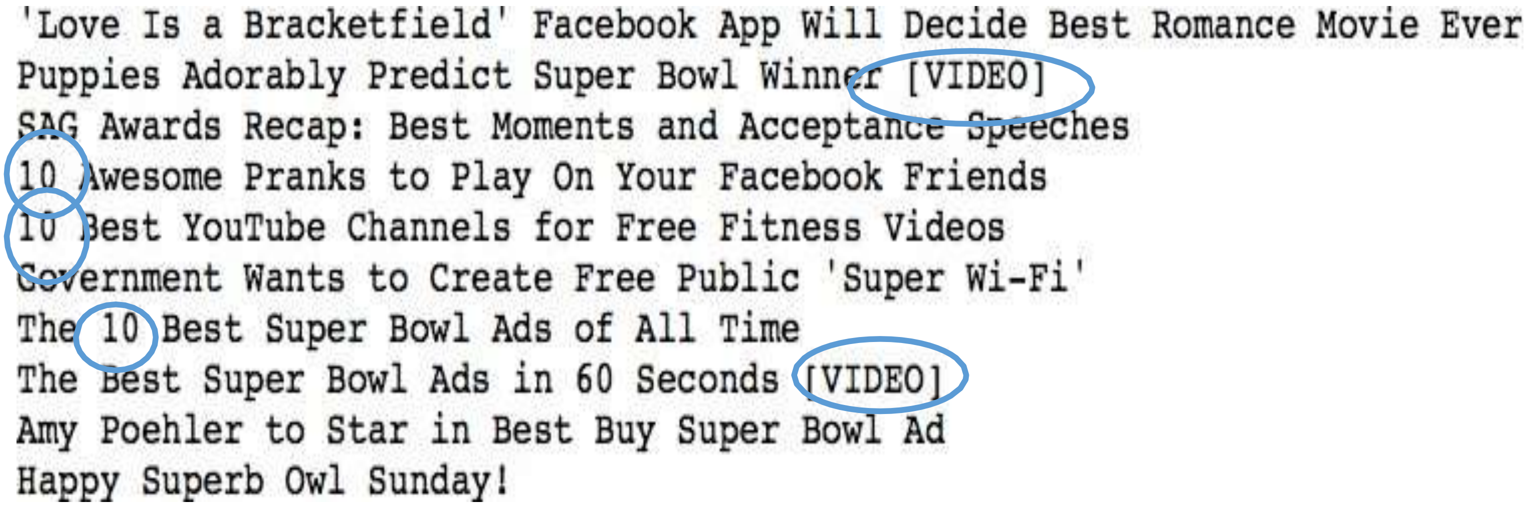
Why entertainment is hot?

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

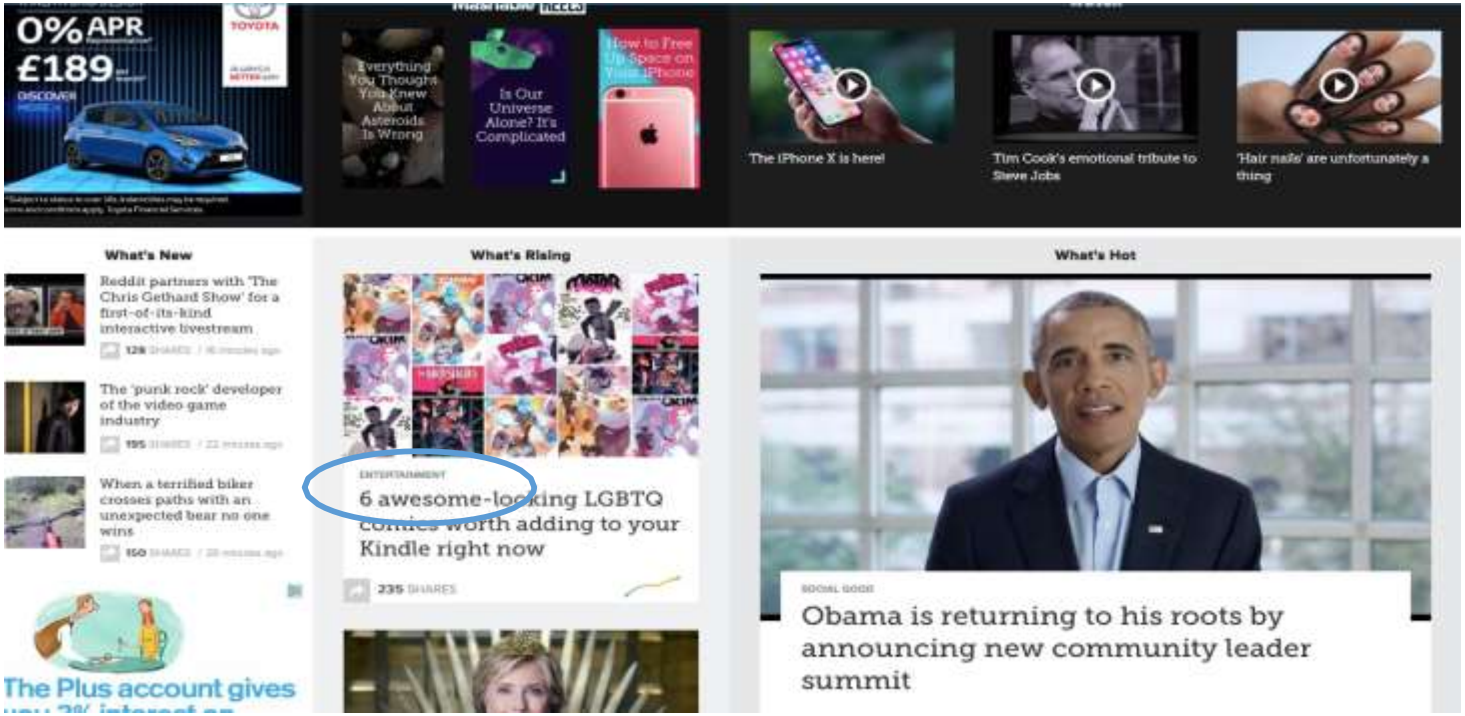
Positive words, strong visual and individual digit in title

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!


0% APR
£189
DISCOVER
TOYOTA
Everything You Thought You Knew About Asteroids Is Wrong
Is Our Universe Alone? It's Complicated
How to Free Up Space on Your iPhone

The iPhone X is here!

Tim Cook's emotional tribute to Steve Jobs

'Hair nails' are unfortunately a thing

What's New

Reddit partners with 'The Chris Gethard Show' for a first-of-its-kind interactive livestream

128 SHARES

The 'punk rock' developer of the video game industry

195 SHARES

When a terrified biker crosses paths with an unexpected bear no one wins

150 SHARES

The Plus account gives

What's Rising

ENTERTAINMENT

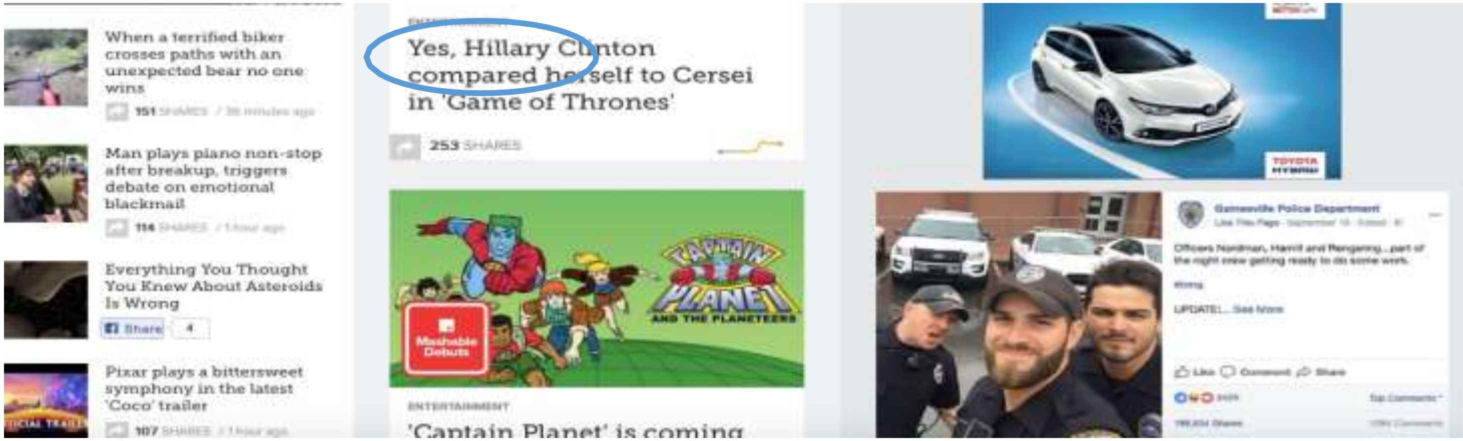
6 awesome-looking LGBTQ comics worth adding to your Kindle right now

235 SHARES

What's Hot

SOCIAL GOOD

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



When a terrified biker crosses paths with an unexpected bear no one wins

151 SHARES

Man plays piano non-stop after breakup, triggers debate on emotional blackmail

114 SHARES

Everything You Thought You Knew About Asteroids Is Wrong

Share 4

Pixar plays a bittersweet symphony in the latest 'Coco' trailer

107 SHARES

ENTERTAINMENT

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

253 SHARES

ENTERTAINMENT

'Captain Planet' is coming

Gainesville Police Department

UPDATE!... See More

Like Comment Share

Title

1. Strong visual
2. Start with a digital numeral
3. Start with positive imperative

SENTIMENT POLARITY

- An article can have both positive and negative polarity which shows the emotions.
- More number of positive polarity many intend towards good articles.
- More number of negative polarity many towards controversial articles.

POSITIVE CORPUS

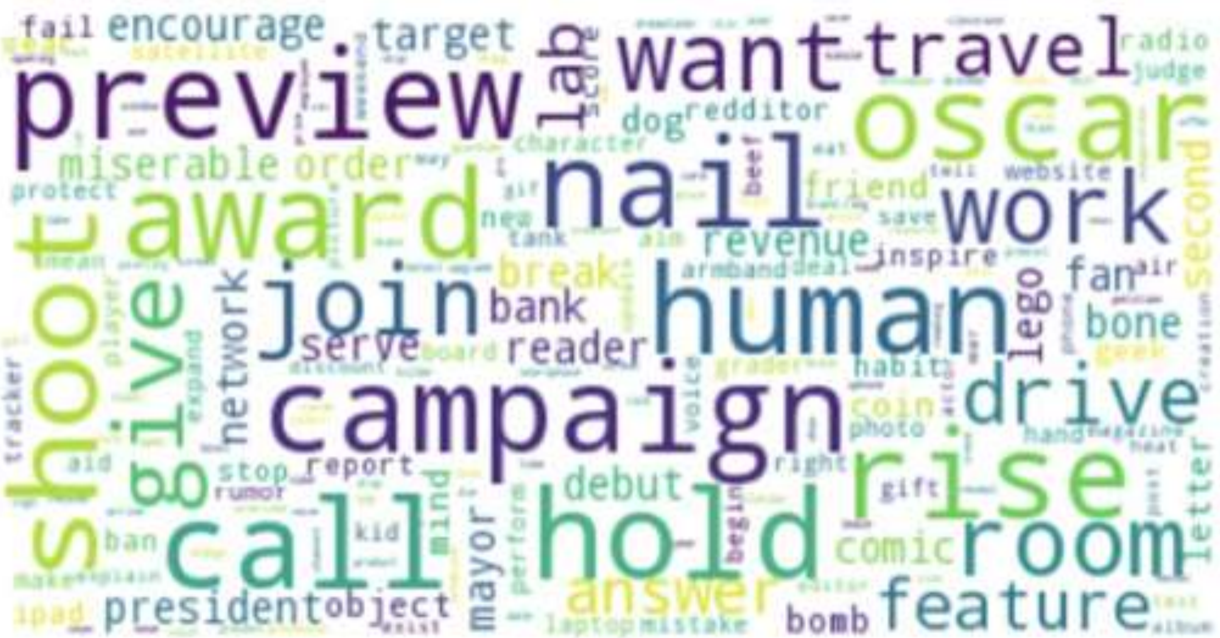
I. “To the world you may be just one person, but to one person you may be the world.”

NEGATIVE CORPOUS

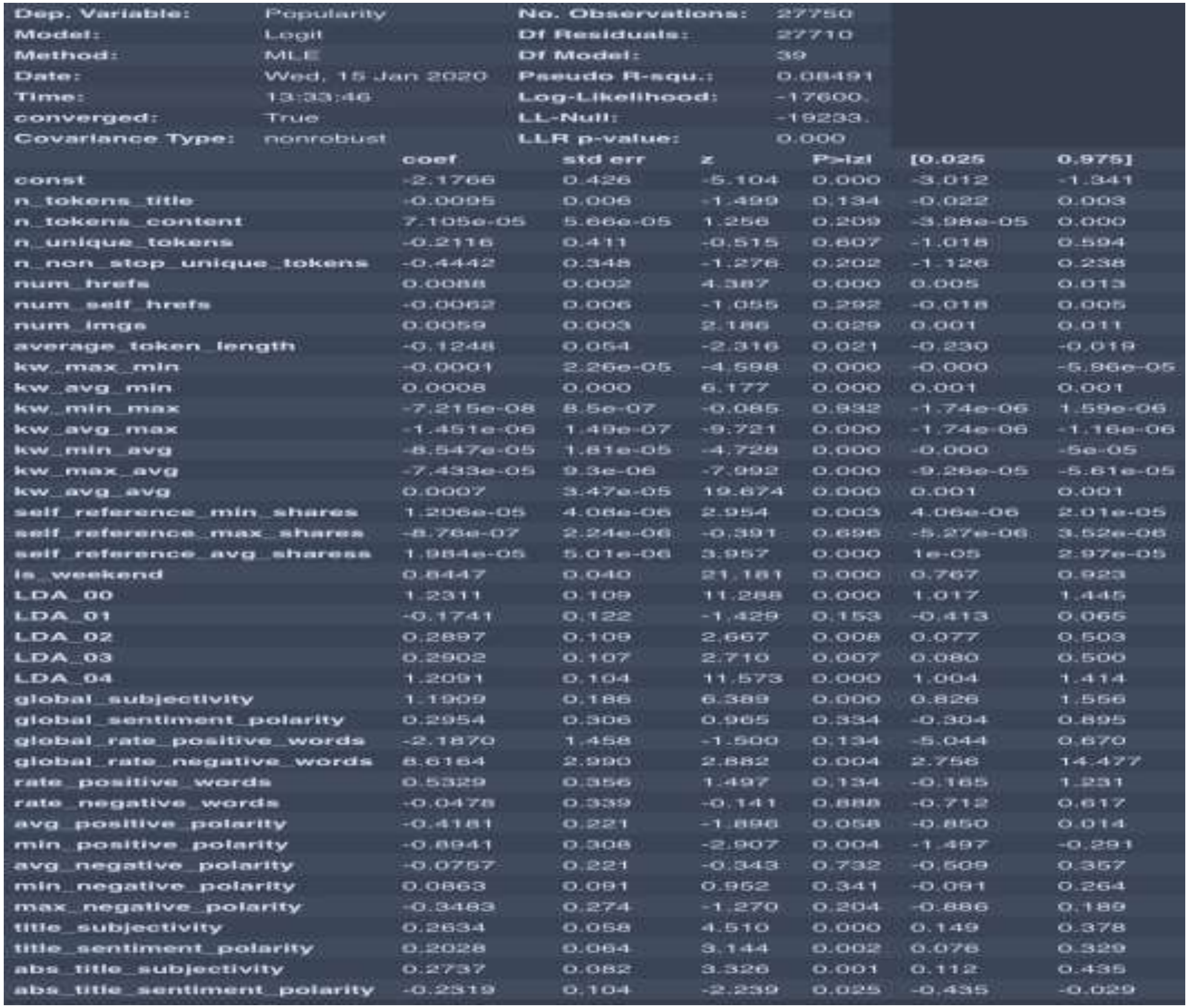
II. “I now believe global warming alarmists are unpatriotic racists knowingly misleading for their own ends. Good night.”

Specific keywords may be more informative

What words were Hot in titles

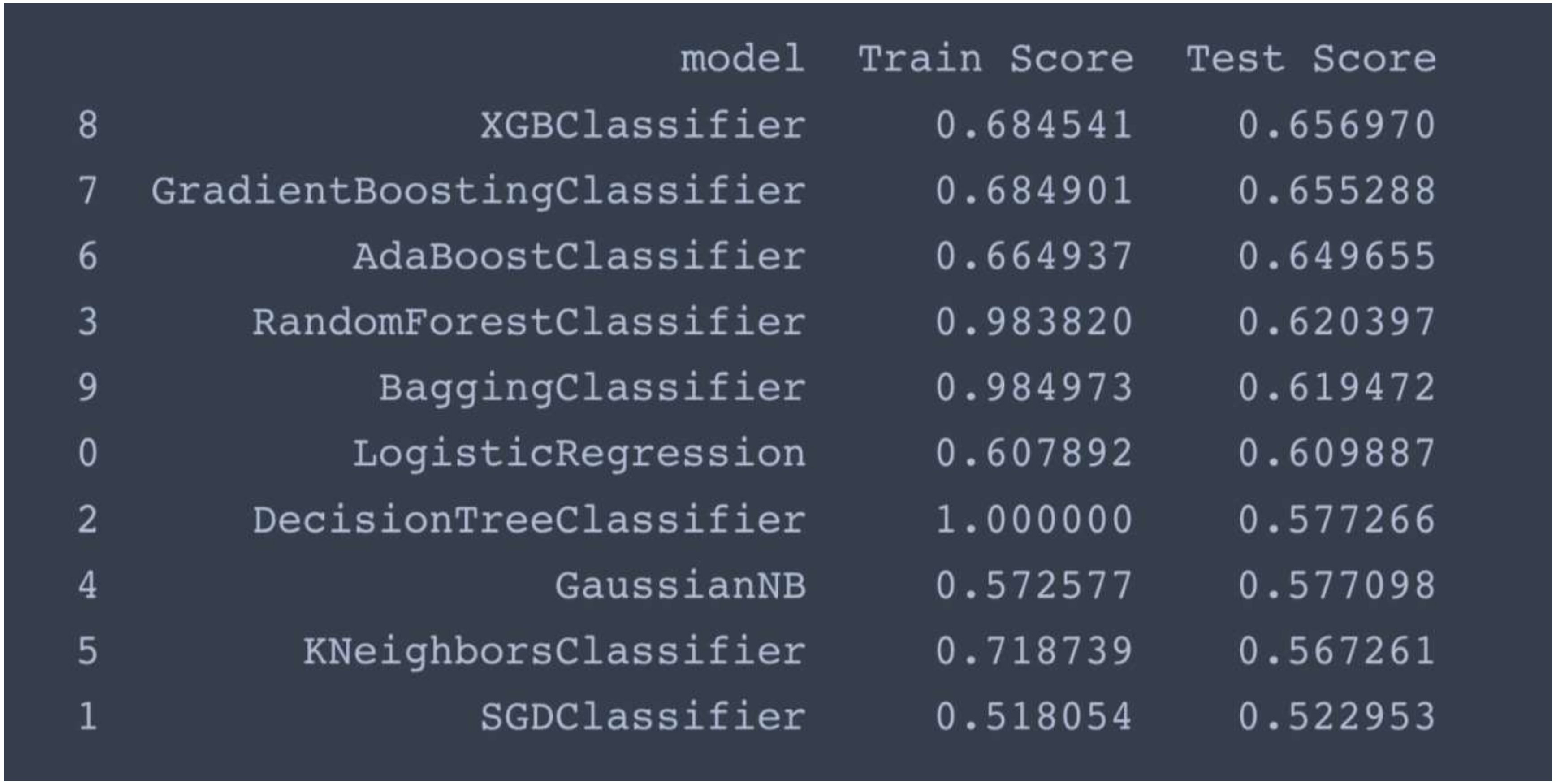


Statistical Model



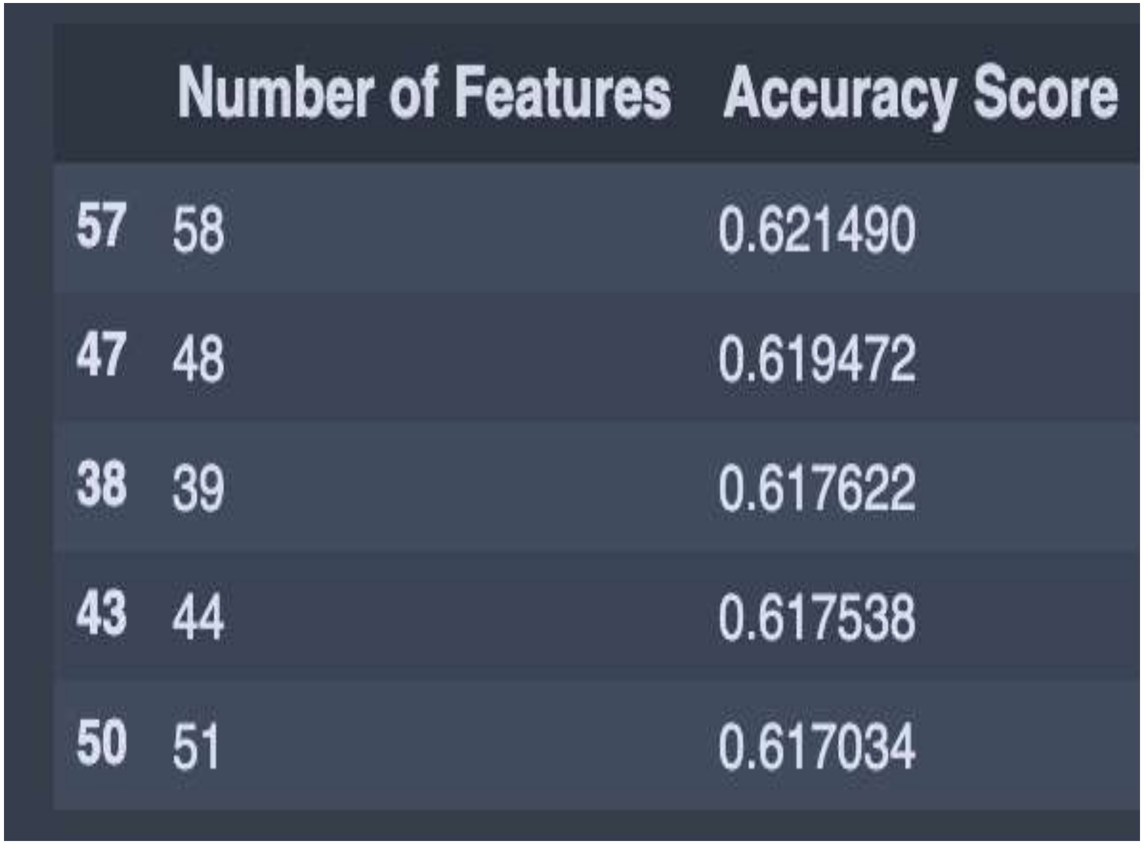
```
Dep. Variable:          Popularity          No. Observations:   27750
Model:                  Logit               Df Residuals:       27710
Method:                 MLE                 Df Model:           39
Date:                   Wed, 15 Jan 2020    Pseudo R-squ.:      0.08491
Time:                   13:33:46            Log-Likelihood:     -17600.
converged:              True                LL-Null:            -19233.
Covariance Type:        nonrobust           LLR p-value:        0.000
                                     coef         std err      z          P>|z|    [0.025       0.975]
const                                -2.1766      0.426        -5.104     0.000    -3.012       -1.341
n_tokens_title                       -0.0095      0.006        -1.499     0.134    -0.022       0.003
n_tokens_content                     7.105e-05    5.66e-05     1.256      0.209    -3.98e-05    0.000
n_unique_tokens                      -0.2116      0.411        -0.515     0.607    -1.018       0.594
n_non_stop_unique_tokens             -0.4442      0.348        -1.276     0.202    -1.126       0.238
num_hrefs                            0.0088       0.002        4.387      0.000    0.005        0.013
num_self_hrefs                       -0.0062      0.006        -1.055     0.292    -0.018       0.005
num_imgs                             0.0059       0.003        2.186      0.029    0.001        0.011
average_token_length                 -0.1248      0.054        -2.316     0.021    -0.230       -0.019
kw_max_min                           -0.0001      2.26e-05     -4.598     0.000    -0.000       -5.86e-05
kw_avg_min                           0.0008       0.000        6.177      0.000    0.001        0.001
kw_min_max                           -7.215e-08   8.5e-07      -0.085     0.932    -1.74e-06    1.59e-06
kw_avg_max                           -1.451e-06   1.49e-07     -9.721     0.000    -1.74e-06    -1.16e-06
kw_min_avg                           -8.547e-05   1.81e-05     -4.728     0.000    -0.000       -5e-05
kw_max_avg                           -7.433e-05   9.3e-06      -7.992     0.000    -9.26e-05    -5.61e-05
kw_avg_avg                           0.0007       3.47e-05     19.674     0.000    0.001        0.001
self_reference_min_shares            1.206e-05    4.06e-06     2.954      0.003    4.06e-06     2.01e-05
self_reference_max_shares            -8.76e-07    2.24e-06     -0.391     0.696    -5.27e-06    3.52e-06
self_reference_avg_sharess           1.984e-05    5.01e-06     3.957      0.000    1e-05        2.97e-05
is_weekend                           0.8447       0.040        21.181     0.000    0.767        0.923
LDA_00                               1.2311       0.109        11.288     0.000    1.017        1.445
LDA_01                               -0.1741      0.122        -1.429     0.153    -0.413       0.065
LDA_02                               0.2897       0.109        2.667      0.008    0.077        0.503
LDA_03                               0.2902       0.107        2.710      0.007    0.080        0.500
LDA_04                               1.2091       0.104        11.573     0.000    1.004        1.414
global_subjectivity                  1.1909       0.186        6.389      0.000    0.826        1.556
global_sentiment_polarity            0.2954       0.306        0.965      0.334    -0.304       0.895
global_rate_positive_words           -2.1870      1.458        -1.500     0.134    -5.044       0.670
global_rate_negative_words           8.6164       2.990        2.882      0.004    2.756        14.477
rate_positive_words                  0.5329       0.356        1.497      0.134    -0.165       1.231
rate_negative_words                  -0.0478      0.339        -0.141     0.888    -0.712       0.617
avg_positive_polarity                -0.4181      0.221        -1.896     0.058    -0.850       0.014
min_positive_polarity                -0.8941      0.308        -2.907     0.004    -1.497       -0.291
avg_negative_polarity                -0.0757      0.221        -0.343     0.732    -0.509       0.357
min_negative_polarity                0.0863       0.091        0.952      0.341    -0.091       0.264
max_negative_polarity                -0.3483      0.274        -1.270     0.204    -0.886       0.189
title_subjectivity                   0.2634       0.058        4.510      0.000    0.149        0.378
title_sentiment_polarity             0.2028       0.064        3.144      0.002    0.076        0.329
abs_title_subjectivity               0.2737       0.082        3.326      0.001    0.112        0.435
abs_title_sentiment_polarity         -0.2319      0.104        -2.239     0.025    -0.435       -0.029
```

Model Deployment with Default Hyper parameters

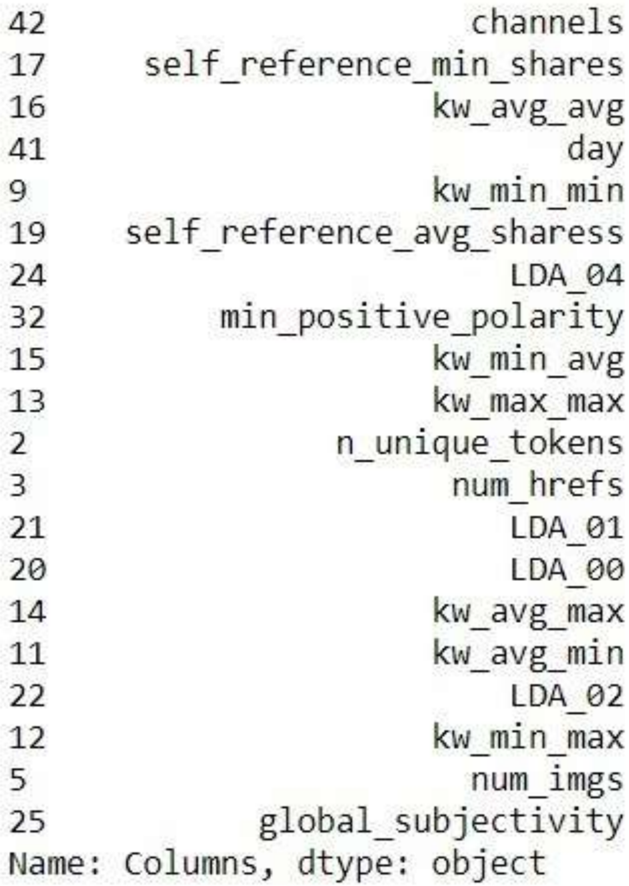


```
                        model  Train Score  Test Score
8               XGBClassifier     0.684541    0.656970
7  GradientBoostingClassifier     0.684901    0.655288
6          AdaBoostClassifier     0.664937    0.649655
3      RandomForestClassifier     0.983820    0.620397
9           BaggingClassifier     0.984973    0.619472
0          LogisticRegression     0.607892    0.609887
2      DecisionTreeClassifier     1.000000    0.577266
4                  GaussianNB     0.572577    0.577098
5        KNeighborsClassifier     0.718739    0.567261
1               SGDClassifier     0.518054    0.522953
```

Feature Selection-RFE

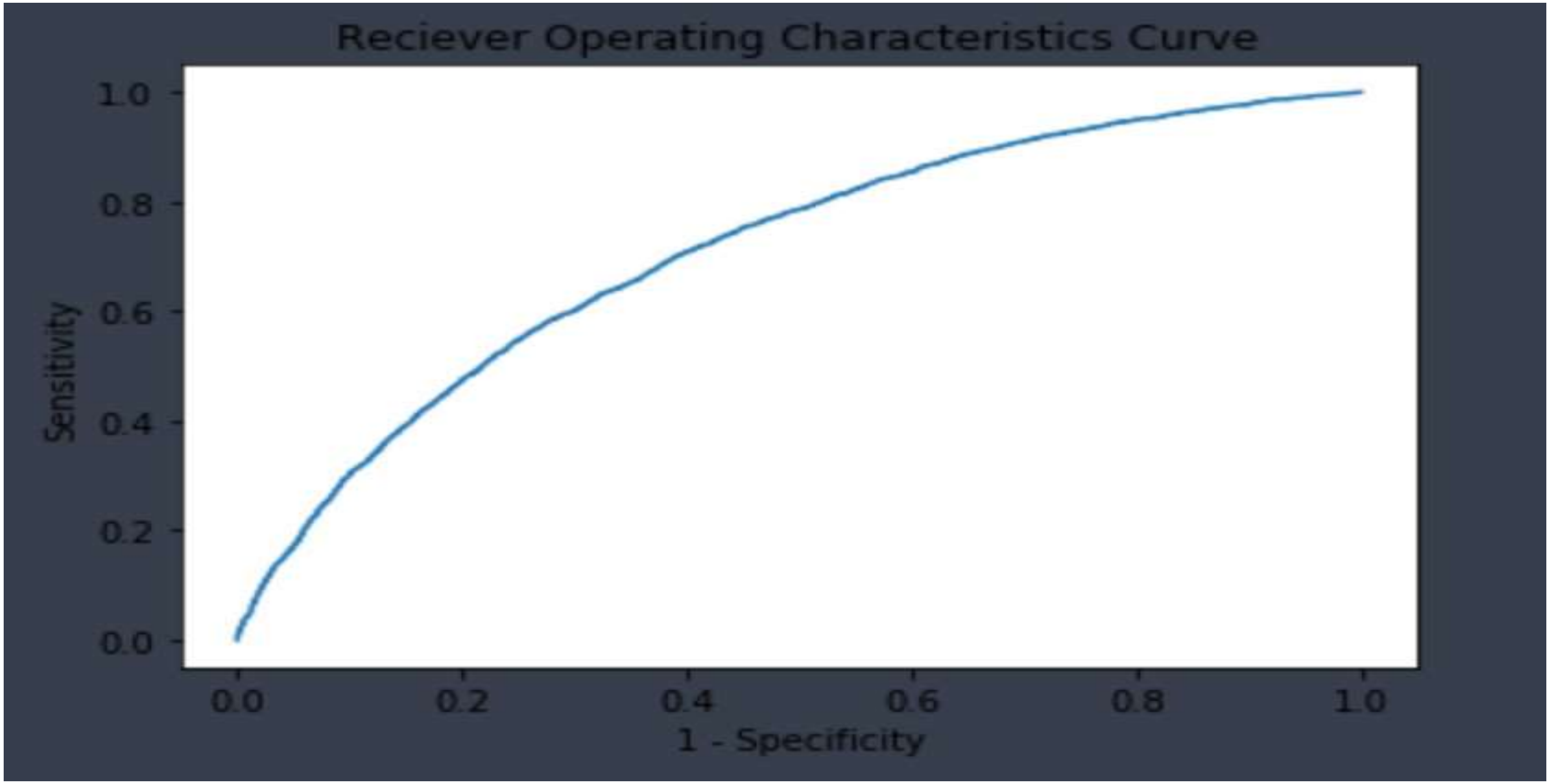


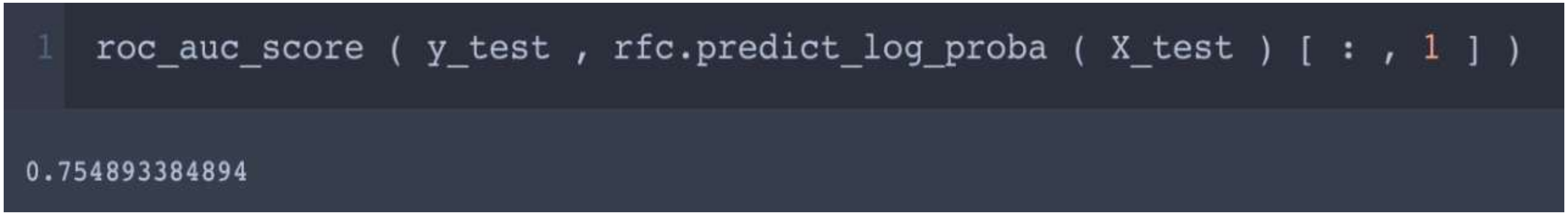
	Number of Features	Accuracy Score
57	58	0.621490
47	48	0.619472
38	39	0.617622
43	44	0.617538
50	51	0.617034



```
42                      channels
17        self_reference_min_shares
16                    kw_avg_avg
41                           day
9                     kw_min_min
19       self_reference_avg_sharess
24                        LDA_04
32           min_positive_polarity
15                    kw_min_avg
13                    kw_max_max
2                n_unique_tokens
3                      num_hrefs
21                        LDA_01
20                        LDA_00
14                    kw_avg_max
11                    kw_avg_min
22                        LDA_02
12                    kw_min_max
5                       num_imgs
25           global_subjectivity
Name: Columns, dtype: object
```

ROC Curve

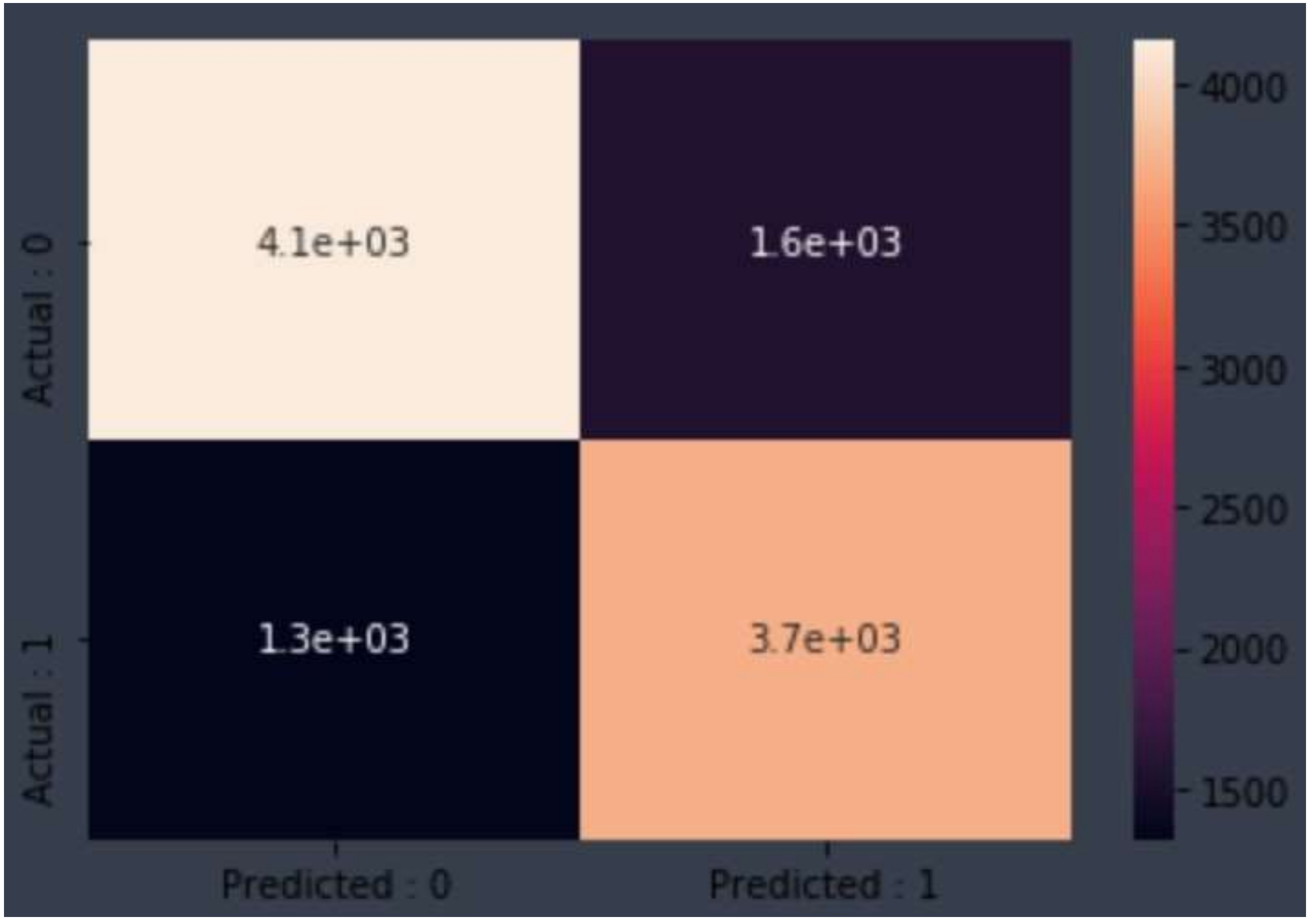


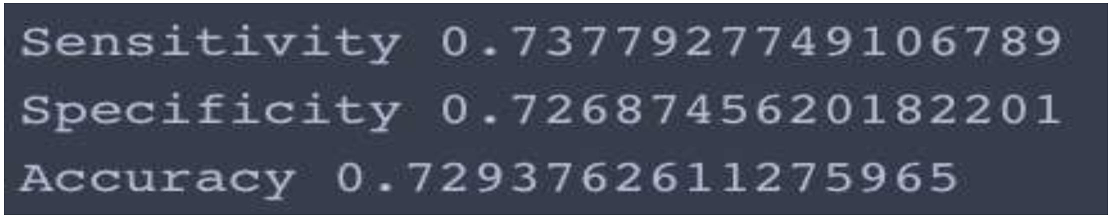


```
roc_auc_score ( y_test , rfc.predict_log_proba ( X_test ) [ : , 1 ] )
```

```
0.754893384894
```

Metrics

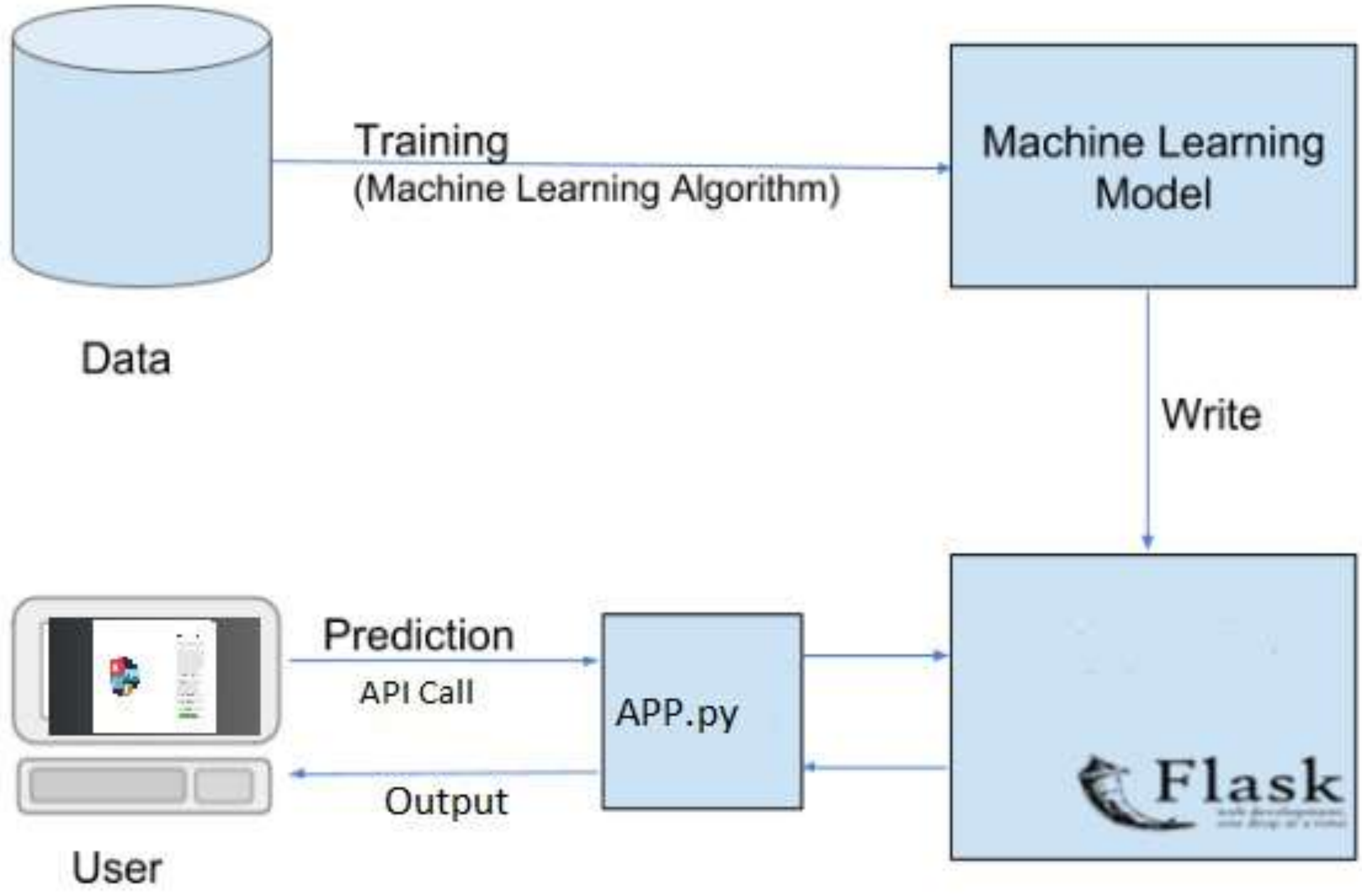




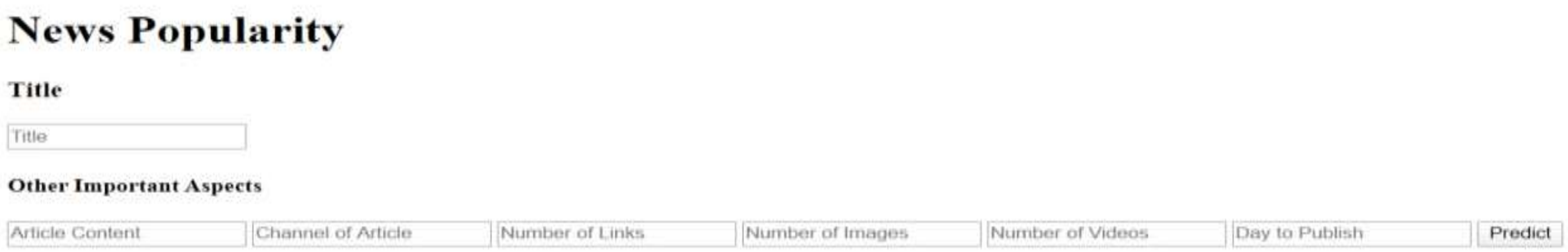
```
Sensitivity 0.7377927749106789
Specificity 0.7268745620182201
Accuracy 0.7293762611275965
```

greatlearning

Architecture



UI: Iteration 1



News Popularity

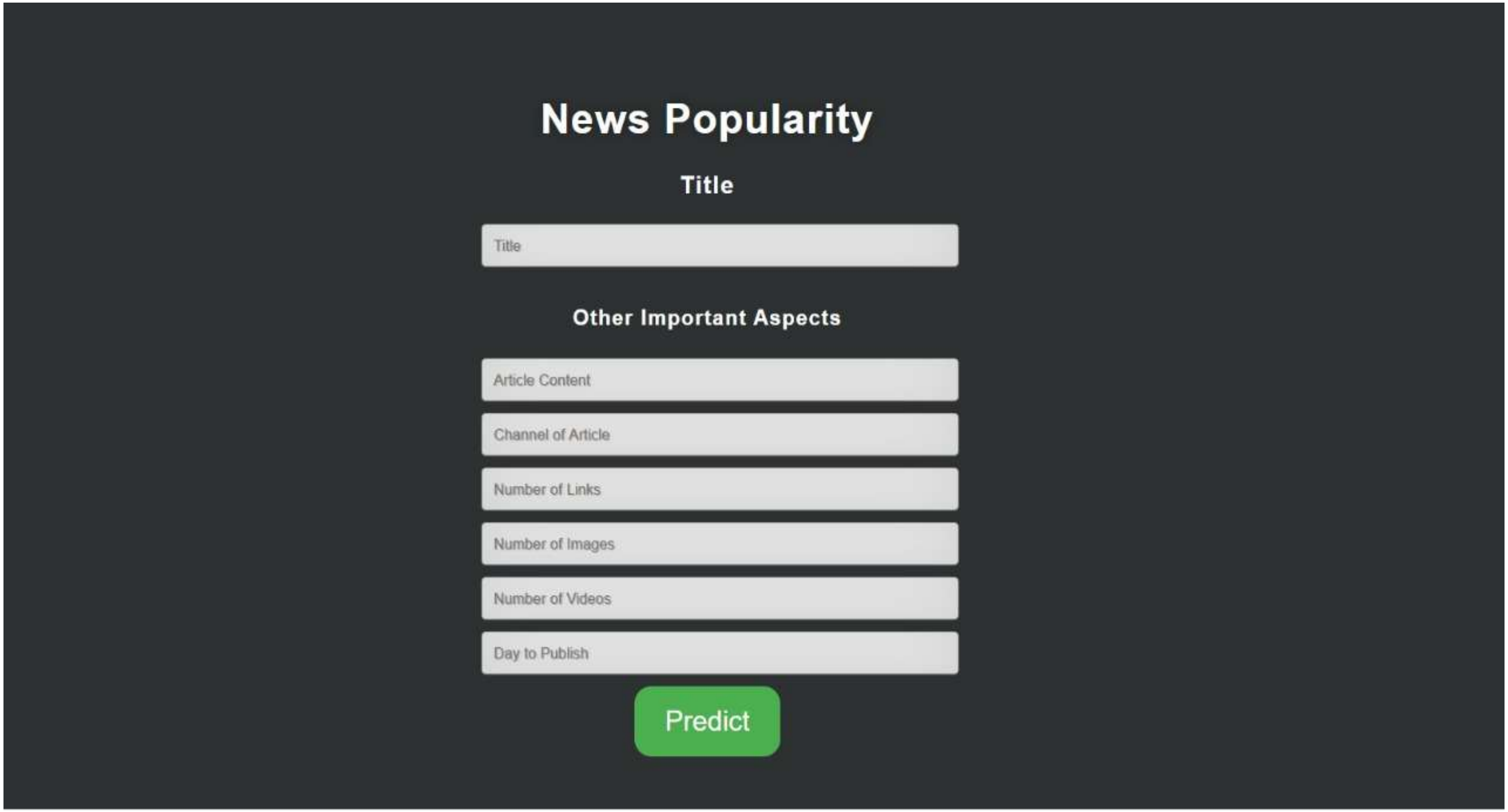
Title

Title

Other Important Aspects

Article Content | Channel of Article | Number of Links | Number of Images | Number of Videos | Day to Publish

UI: Iteration 2



News Popularity

Title

Title

Other Important Aspects

Article Content

Channel of Article

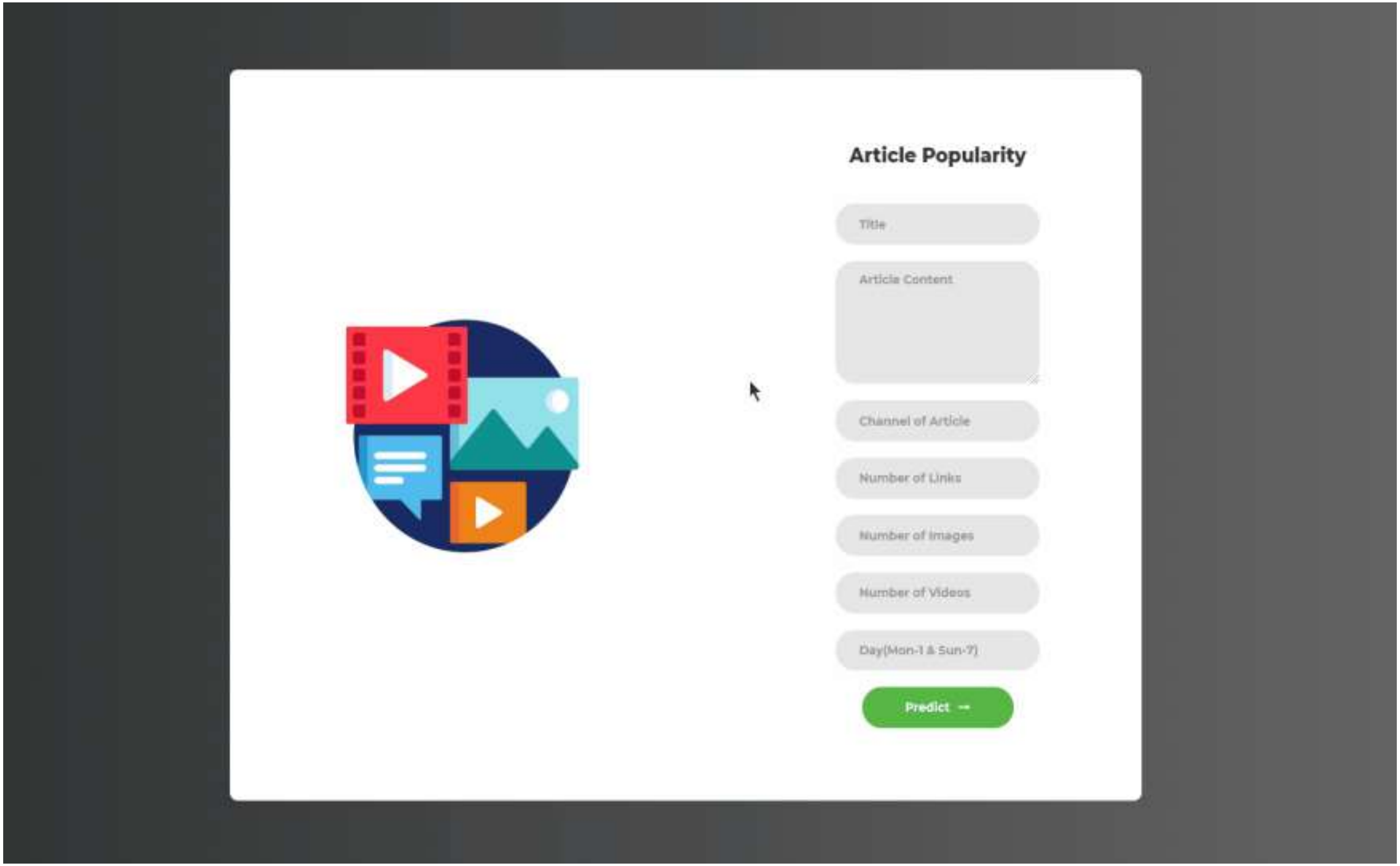
Number of Links

Number of Images

Number of Videos

Day to Publish

User Interface



API as a service

{url}/predict

1. Different UI
2. No UI

Limitations:

- Platform Dependent
- Accuracy VS Inference-time trade off
- Lack of feature details
  - 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

Future Enhancements:

- Content assistance by providing suggestions to the user.
- We can use more complex models at the cost of less interpretability.
- The next iteration – URL Extraction

Please take out your Phones!!!

kahoot.it

1 2 3 4 5 6 7 8 9

3x

$200

Google Cloud

THANK YOU!

ngiyabonga
شكراً جزيلاً
danke
謝謝
teşekkür ederim
спасибо
dank je
thank you
gracias
mochchakkeram
dziekuje
go raibh maith agat
obrigado
sukriya
kop khun krap
grazie
arigato
takk
dakujem
мерси
terima kasih
감사합니다
merci