Online News Popularity

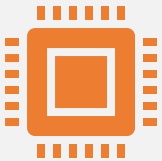
DIA1

Valentin ABOU-OBEIDA

Luca RINGUET

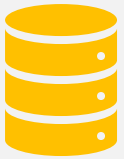
Clarisse SACRE

Summary



I – Data Set Information

II - Data visualization



III - Modeling

I - Data Set Information

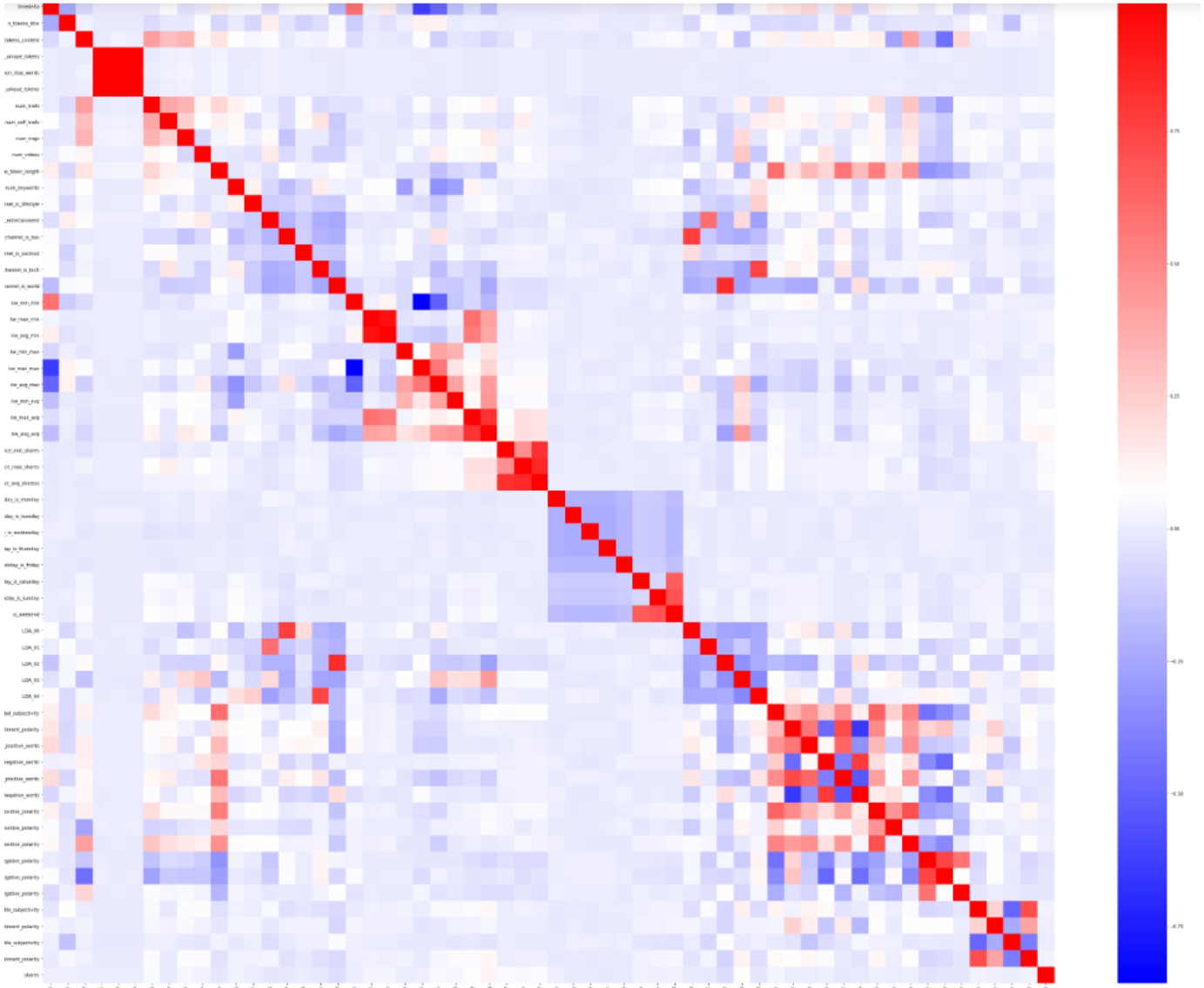
Articles from Mashable

January 8, 2015

Period of 2 years

II - Data visualization

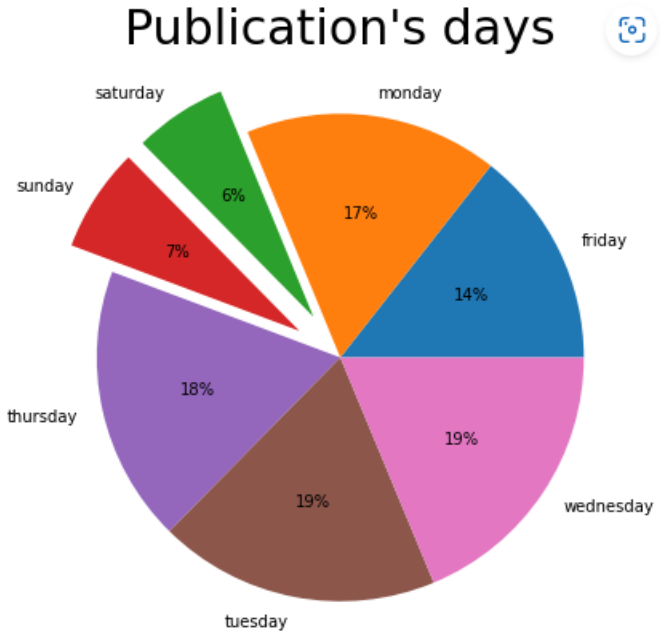
- Corrélation Matrix

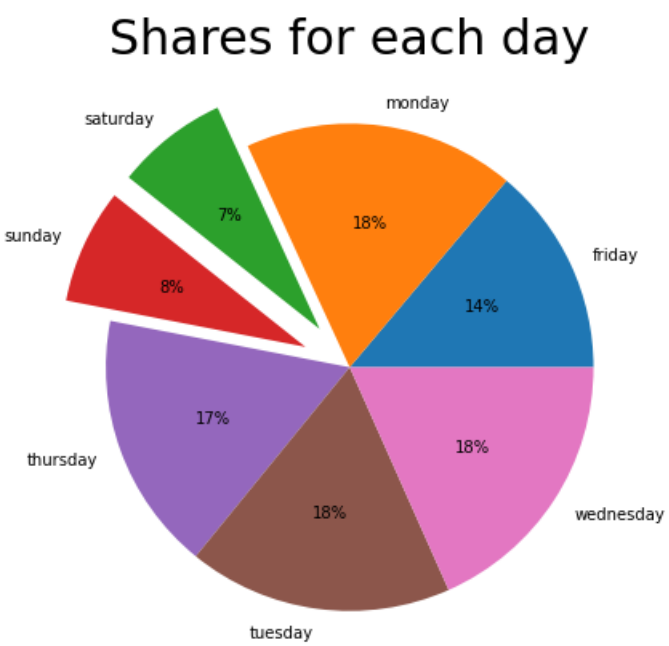


II - Data visualization

Studies around articles' type and day

Ranking days of publications and shares

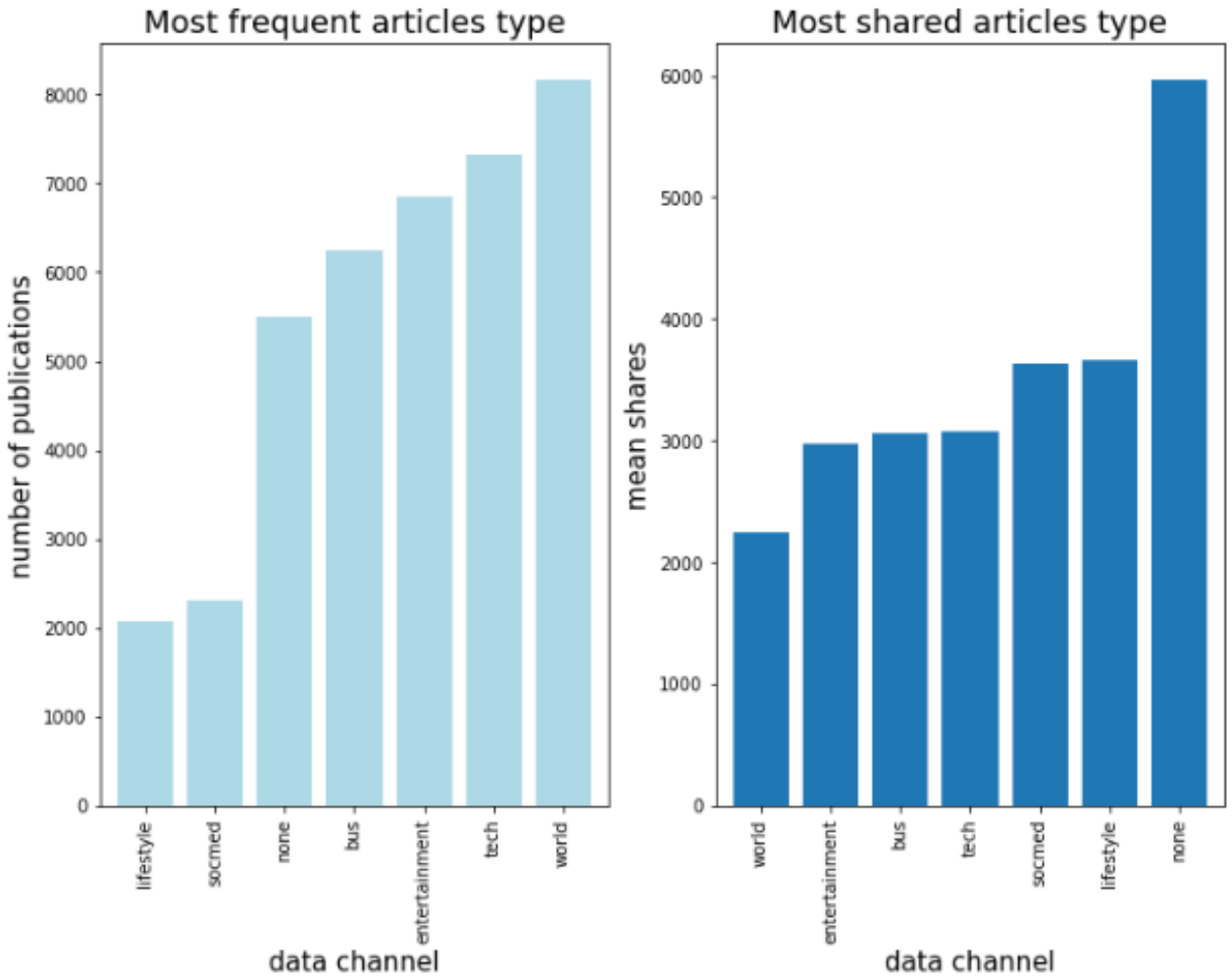




II - Data visualization

Studies around articles' type and day

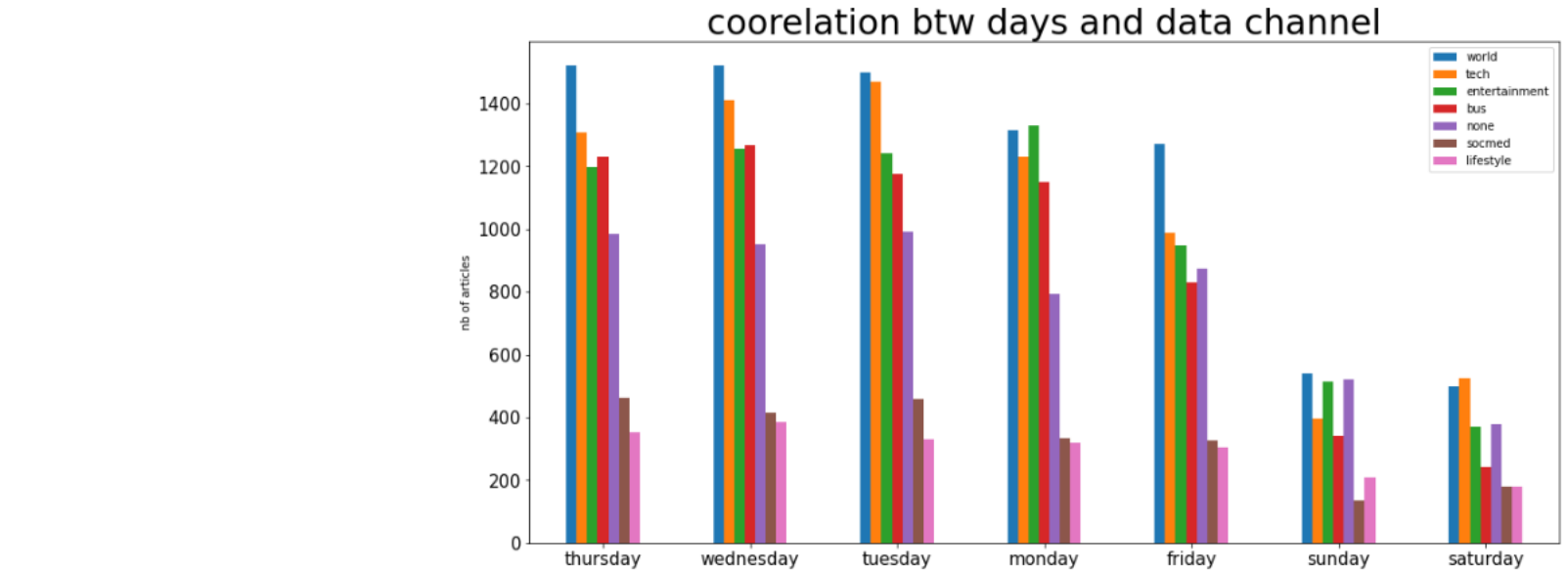
Most frequent/shares articles' type

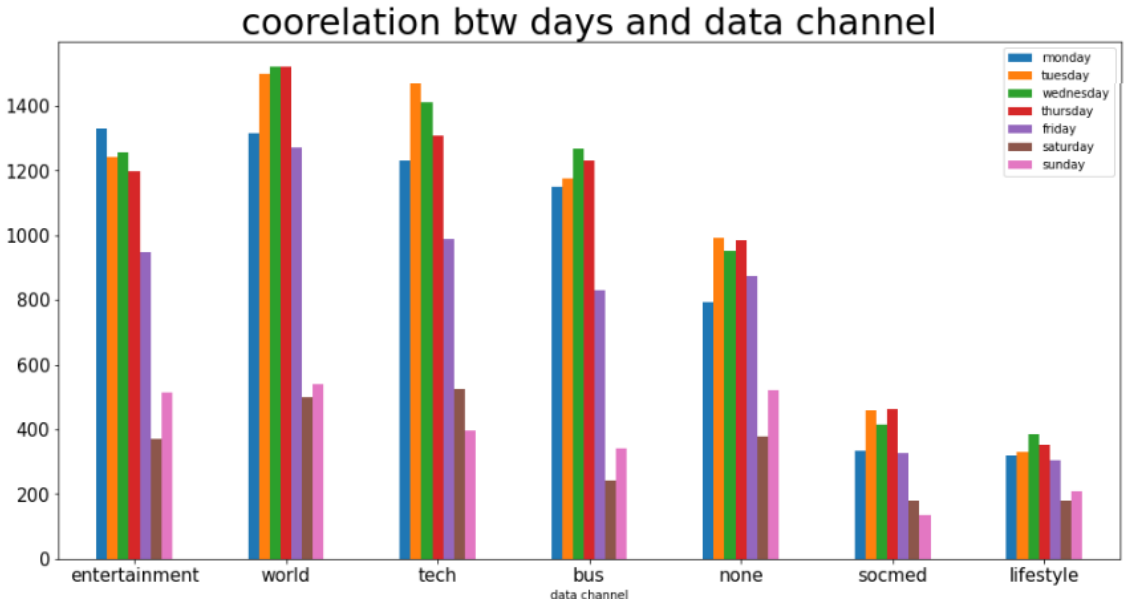


II - Data visualization

Studies around articles' type and day

Most frequent/shares articles' type





II - Data visualization

Studies around articles' type and day

Top positives articles

title	
iranian dog shelters are rare	socmed
2013 oscar predictions polls	entertainment
apple stock 2	tech
google ideas comics	none
fiverr funding	none
hyperloop daryl oster	none
internet art comic	none
superheroic letdown	none
tomatina festival tomato food fight	socmed
free diy projects	none

iranian dog shelters are rare	socmed
2013 oscar predictions polls	entertainment
apple stock 2	tech
tomatina festival tomato food fight	socmed
apple new privacy policy	bus
act like australian	socmed
american apparel standard general deal	socmed
minecraft earnings brief	entertainment
nike fuelband ces 2014	bus
instagram video uploads brands	tech

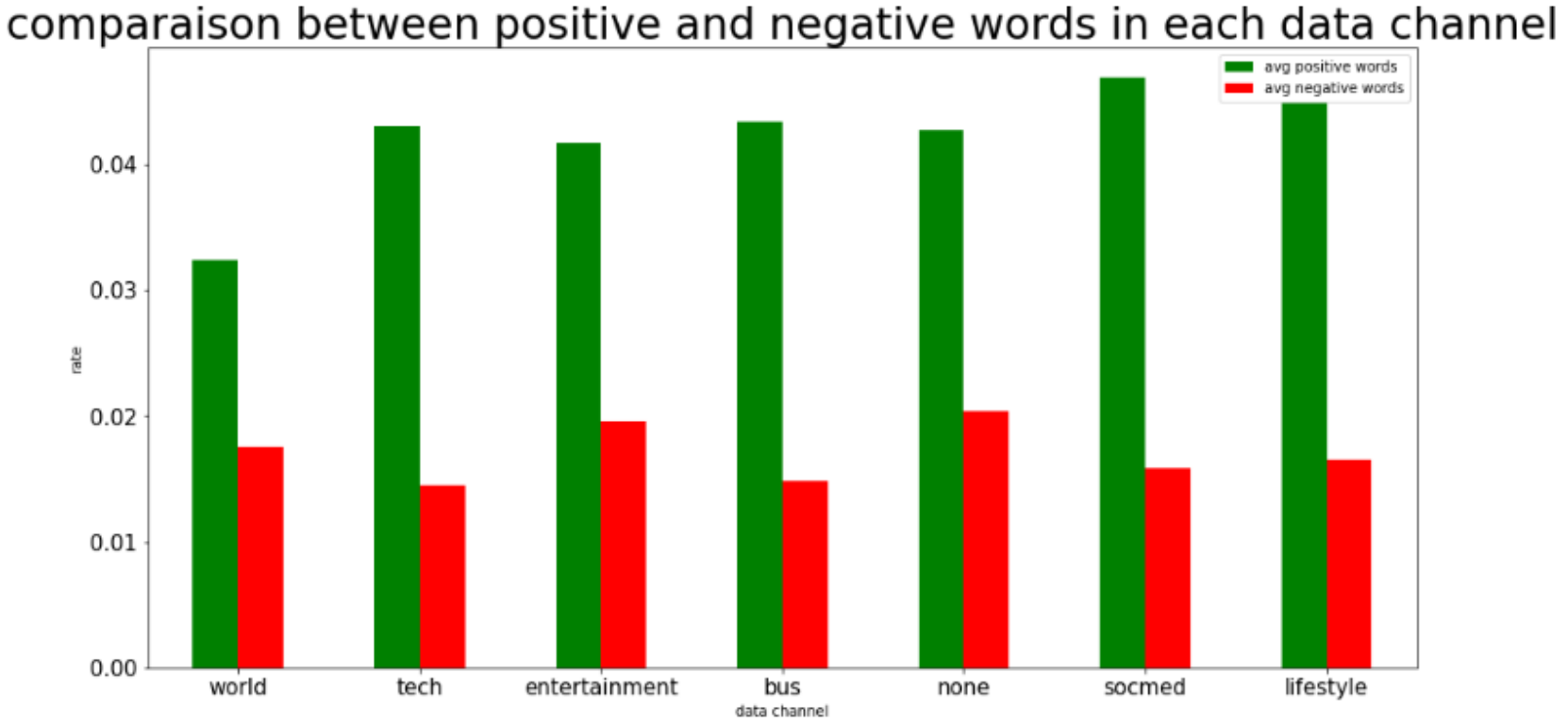
Top negatives articles

title	
twitter reacts to apple beats acquisition	none
black cats bad for photos	none
fuck yeah tumblrs	socmed
jobs movie josh gad interview	none
movie vine challenge roundup	none
facebook unpublished posts	none
fort hood shooter	none
uber missing tips	socmed
ukraine russia war	none
lg retro tv	none

fuck yeah tumblrs	socmed
uber missing tips	socmed
george r r martin red wedding	entertainment
snl drunk uncle peter dinklage	entertainment
game of thrones reactions video	entertainment
kirk douglas quotes	entertainment
dead cellphone could deny boarding	tech
dancing with the stars abc comedy	entertainment
hulu april fools day	entertainment
red vine world record	entertainment

II - Data visualization

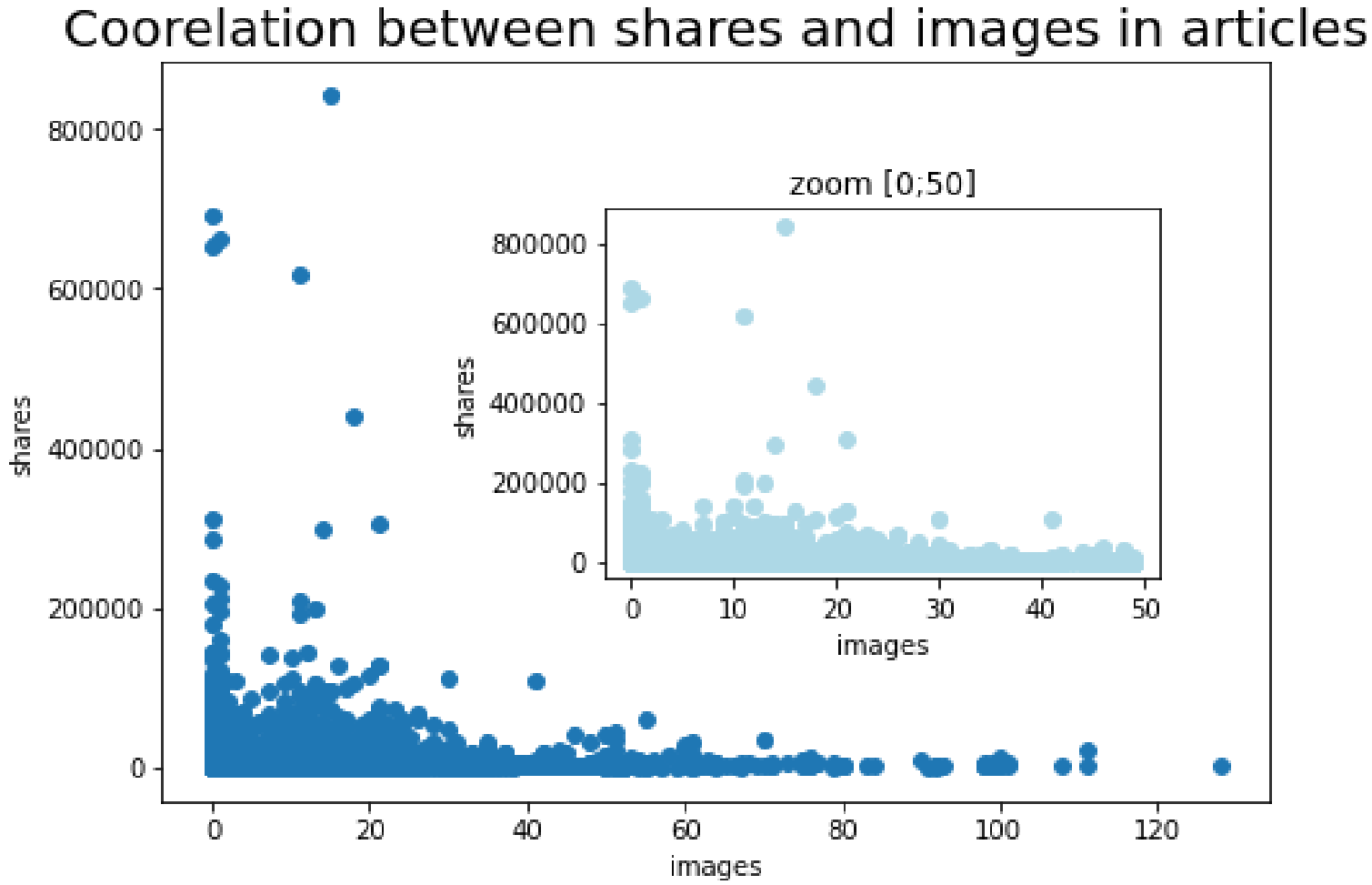
Studies around articles' type and day




```
average positive words in all articles :  0.04
average negative words in all articles :  0.02
```

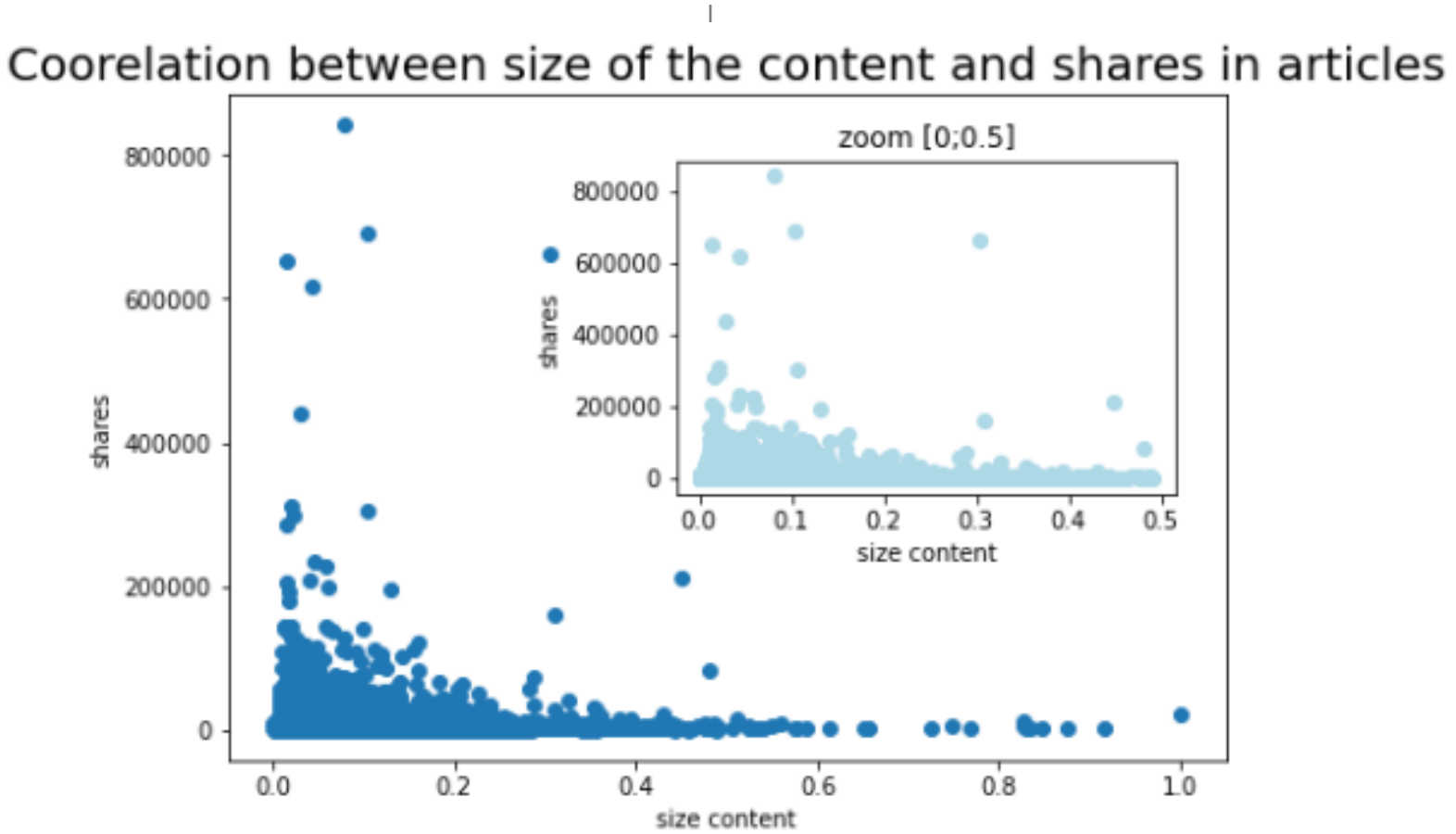
II - Data visualization

Studies around shares



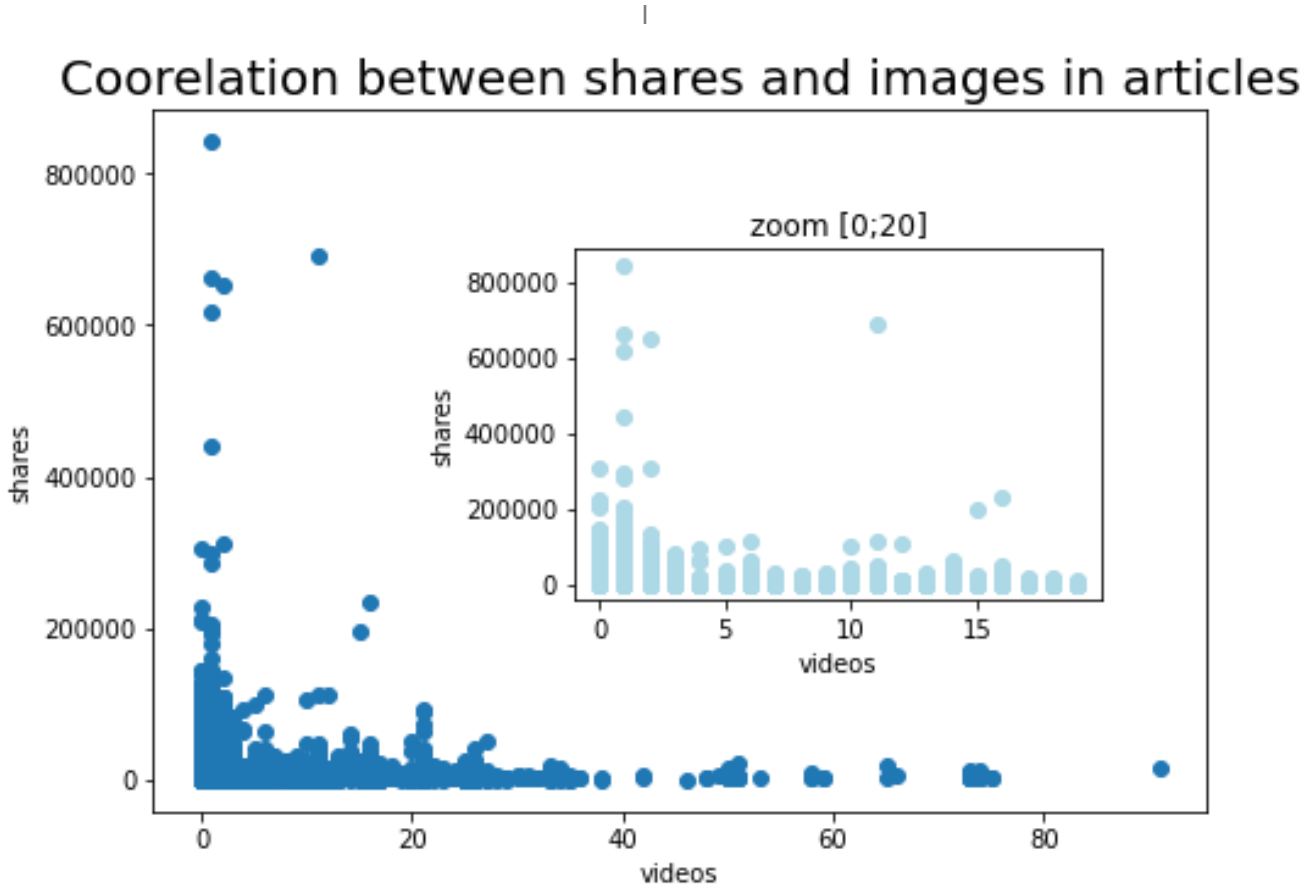
II - Data visualization

Studies around shares



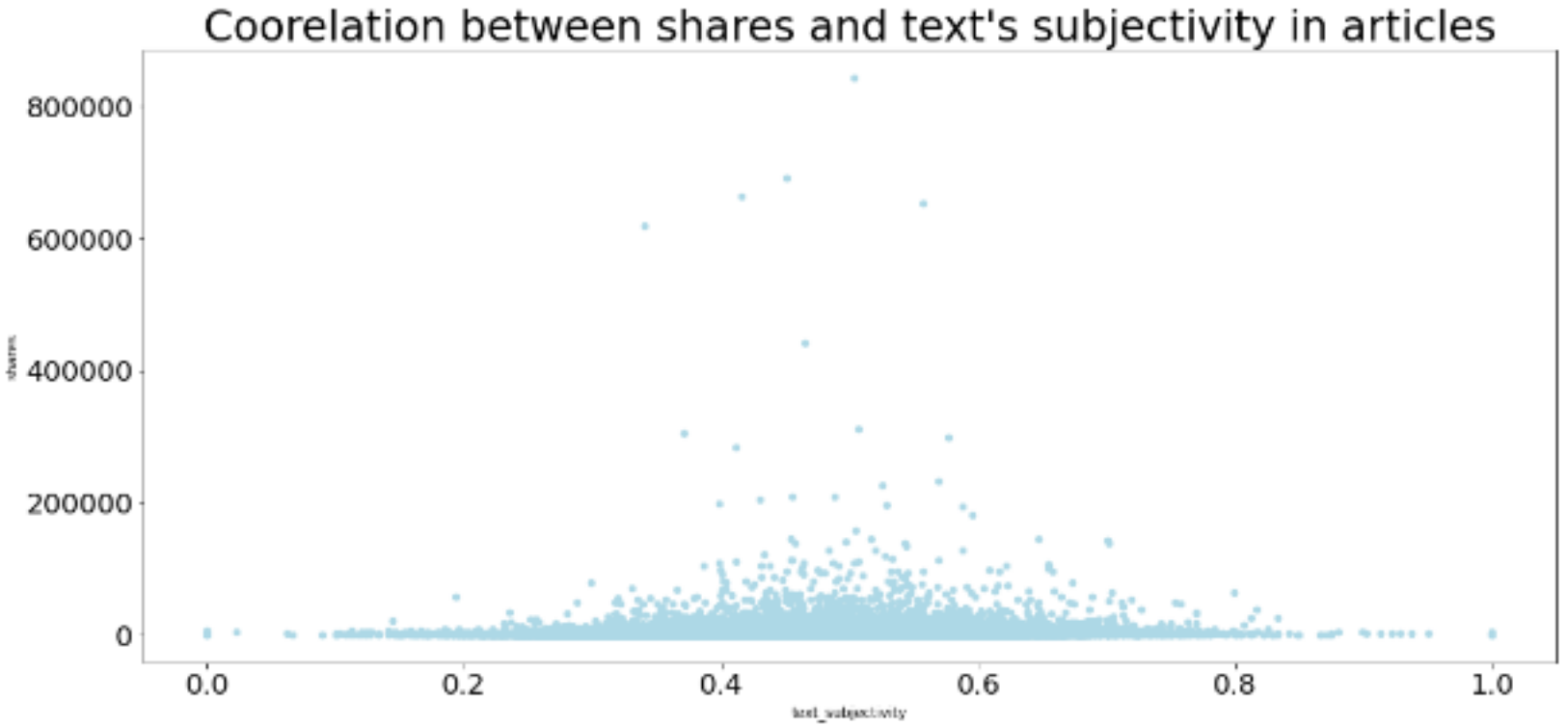
II - Data visualization

Studies around shares



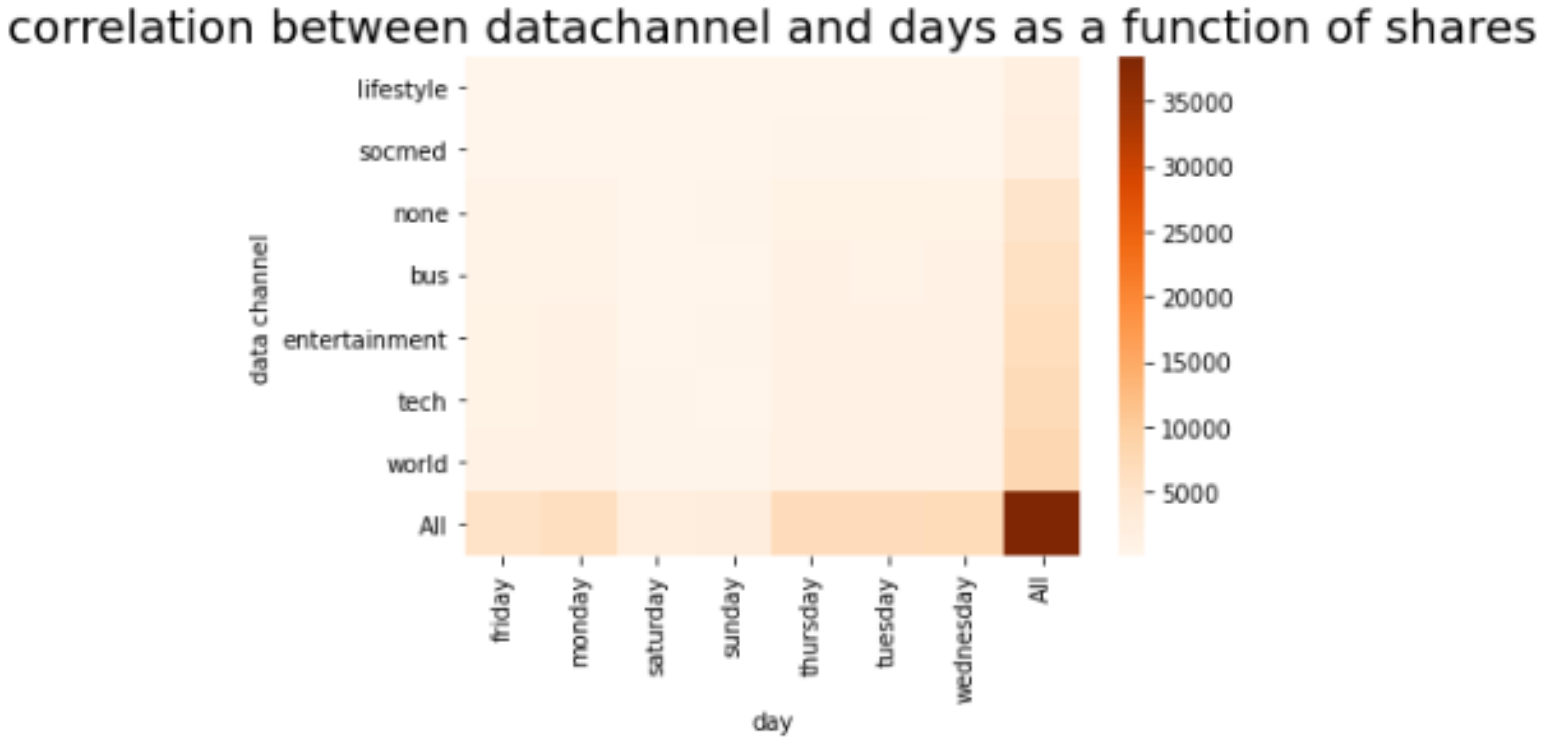
II - Data visualization

Studies around shares



II - Data visualization

Studies around shares



II - Modeling

LinearRegression

```
* Forecasting Metrics Summary *
RMSE: 4394.145375753702
MAE: 2266.425498285506
R²: 0.054
```

Ridge Regression

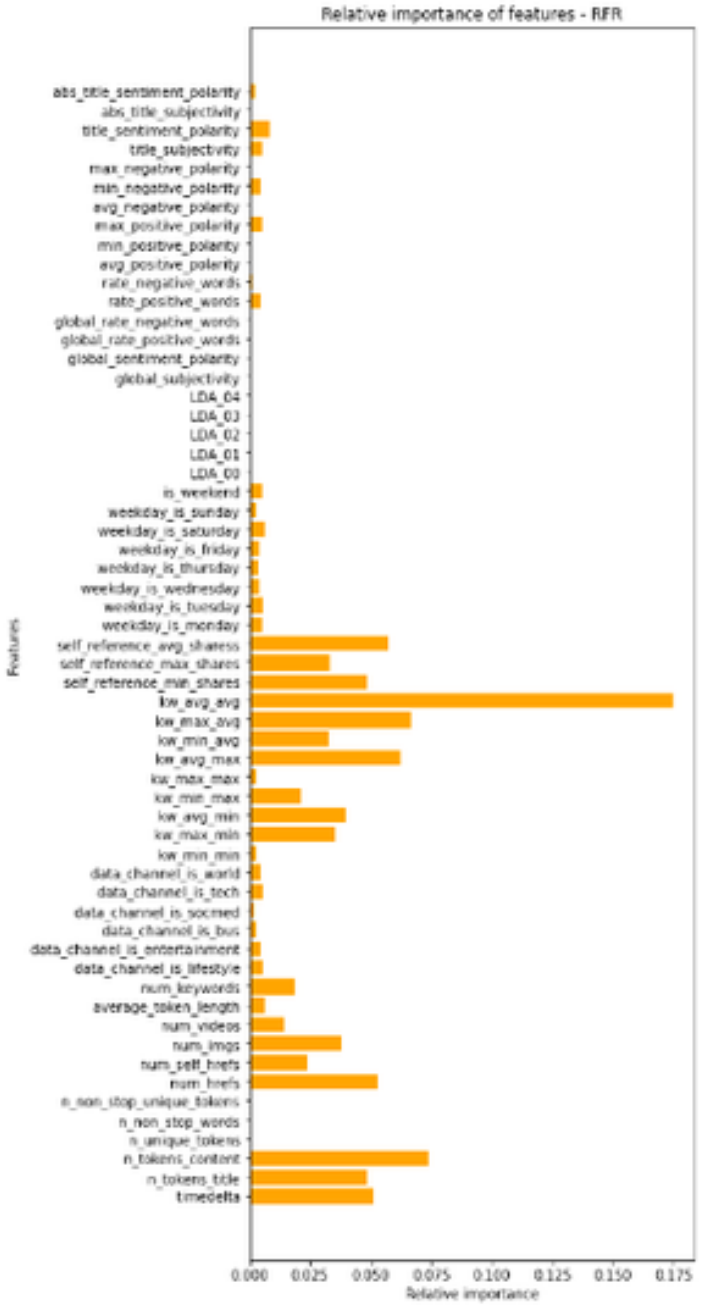
```
* Forecasting Metrics Summary *
RMSE: 4394.1519317863895
MAE: 2266.4499920596904
R²: 0.054
```

Random Forest regressor

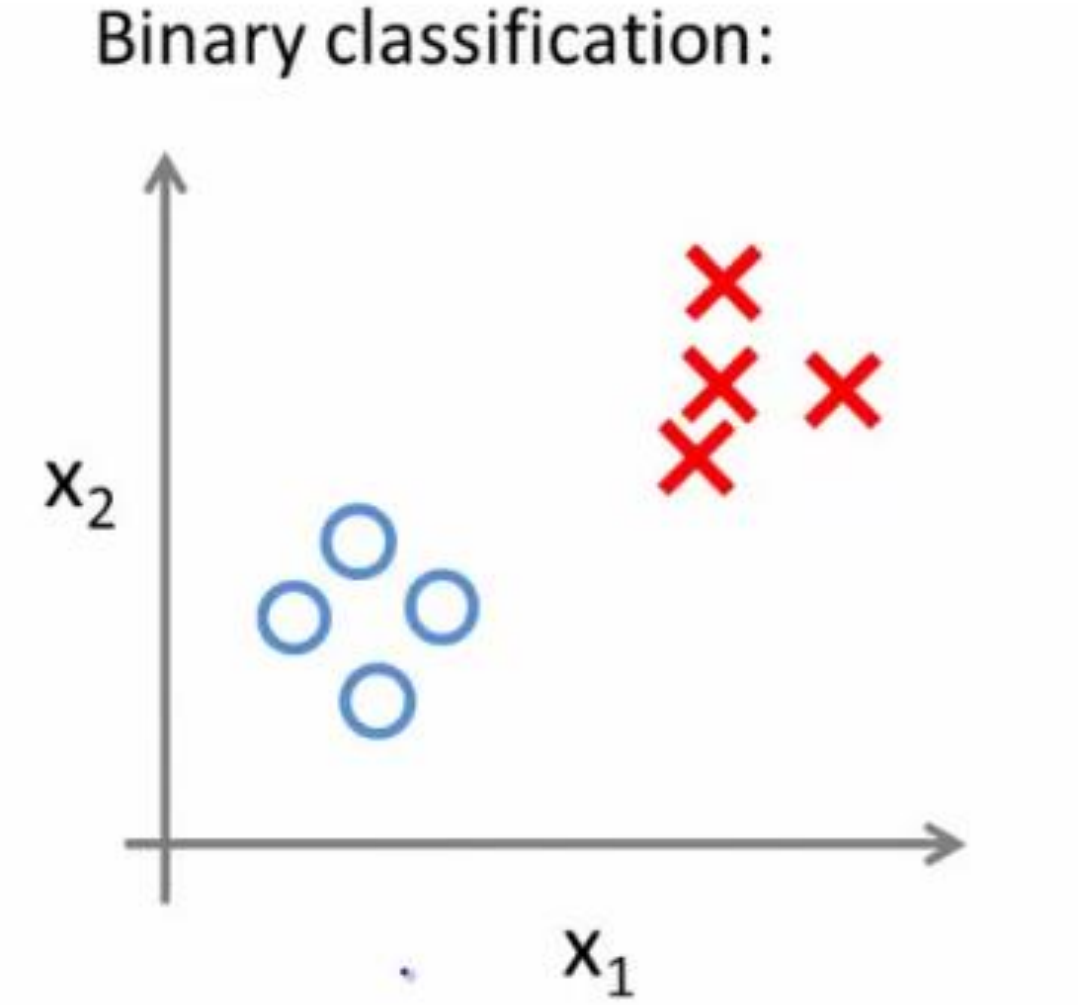
```
* Forecasting Metrics Summary *
RMSE: 4383.557225540685
MAE: 2254.1148704569196
R²: 0.058
```

Optimization random forest regressor

```
* Forecasting Metrics Summary *
RMSE: 4411.190616433405
MAE: 2277.90517203631
R²: 0.046
```



Classification , a better option



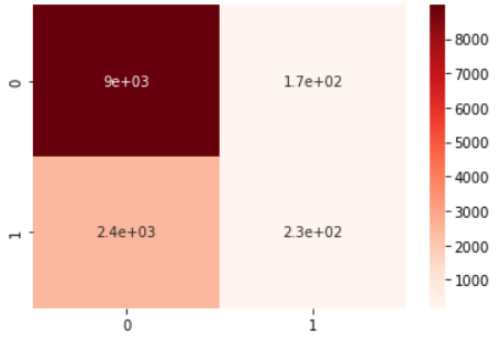
"1" : the article is popular, "0" : the article is unpopular

	precision	recall	f1-score	support
0	0.7738	0.9999	0.8724	9153
1	0.8000	0.0015	0.0030	2680
accuracy			0.7738	11833
macro avg	0.7869	0.5007	0.4377	11833
weighted avg	0.7797	0.7738	0.6755	11833

Sensibility probability

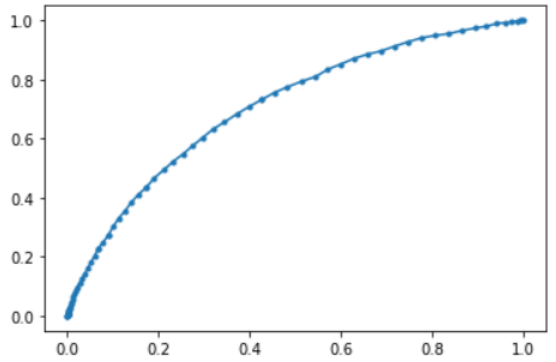
	precision	recall	f1-score	support
0	0.7859	0.9811	0.8727	9153
1	0.5749	0.0873	0.1516	2680
accuracy			0.7787	11833
macro avg	0.6804	0.5342	0.5122	11833
weighted avg	0.7381	0.7787	0.7094	11833

Correlation Matrix



RandomForest : AUROC = 0.709

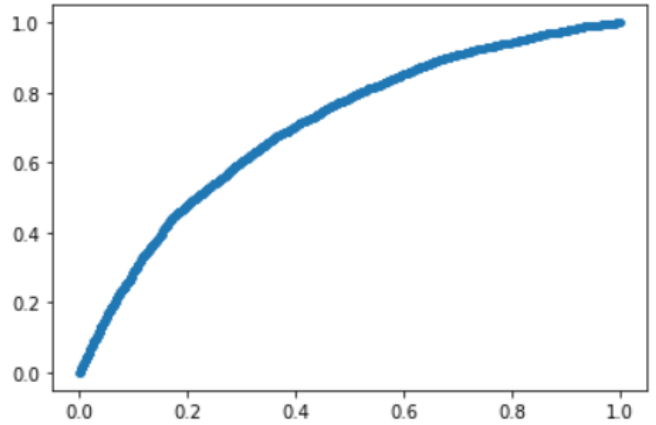
[<matplotlib.lines.Line2D at 0x26366a9cd90>]



RandomForestClassifier

Second Classification, Adaptive Boost Classifier

```
AdaBoost : AUROC = 0.707
[<matplotlib.lines.Line2D at 0x26366b08e80>]
```

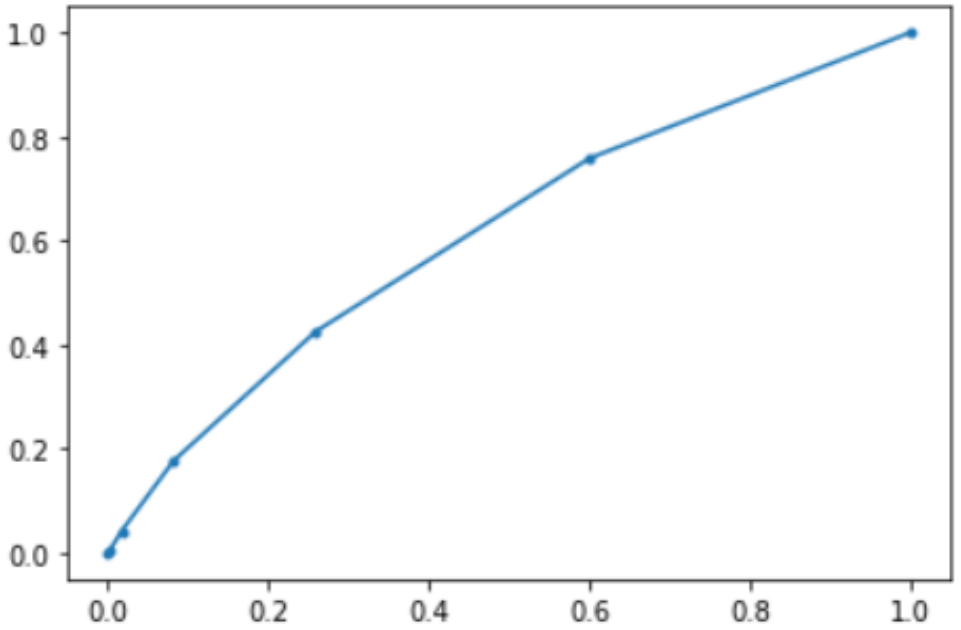


	precision	recall	f1-score	support
0	0.7860	0.9697	0.8682	9153
1	0.4870	0.0981	0.1634	2680
accuracy			0.7723	11833
macro avg	0.6365	0.5339	0.5158	11833
weighted avg	0.7183	0.7723	0.7086	11833

First model scare, K-Nearest Neighbours :

KNN : AUROC = 0.614

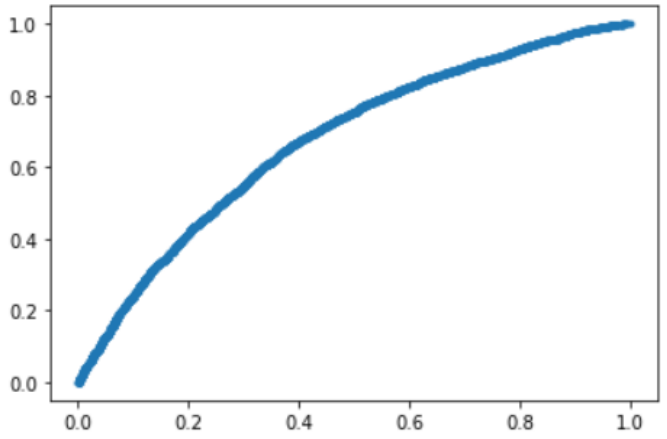
[<matplotlib.lines.Line2D at 0x13f1111a8e0>]



	precision	recall	f1-score	support
0	0.7921	0.9180	0.8504	9153
1	0.3874	0.1772	0.2432	2680
accuracy			0.7502	11833
macro avg	0.5898	0.5476	0.5468	11833
weighted avg	0.7005	0.7502	0.7129	11833

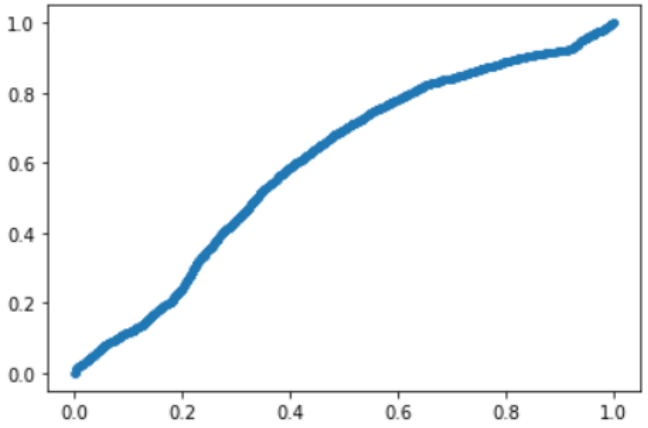
```
KNN : AUROC = 0.676

[<matplotlib.lines.Line2D at 0x13f1862aaf0>]
```




```
KNN : AUROC = 0.603

[<matplotlib.lines.Line2D at 0x13f1159b850>]
```



	precision	recall	f1-score	support
0	0.7756	0.9743	0.8637	9153
1	0.2985	0.0373	0.0663	2680
accuracy			0.7621	11833
macro avg	0.5371	0.5058	0.4650	11833
weighted avg	0.6676	0.7621	0.6831	11833

```
Gauss_acc=accuracy_score(y_test, y_pred_gauss)
print("Accuracy Gauss :",Gauss_acc)
```

```
Accuracy Gauss : 0.762105974816192
```

```
print(classification_report(y_test, y_pred_bern, digits
```

	precision	recall	f1-score	support
0	0.8245	0.7986	0.8114	9153
1	0.3788	0.4194	0.3981	2680
accuracy			0.7128	11833
macro avg	0.6017	0.6090	0.6047	11833
weighted avg	0.7236	0.7128	0.7178	11833

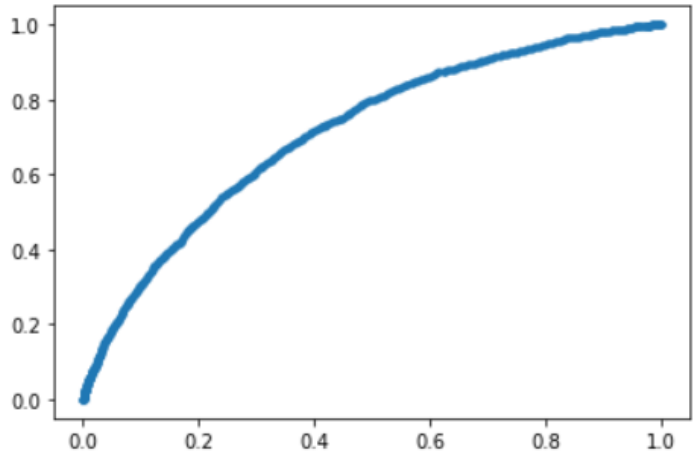
Second model scale, Naive Bayes (Gaussian and Bernoulli):

Comparing and printing all our models results without optimization:

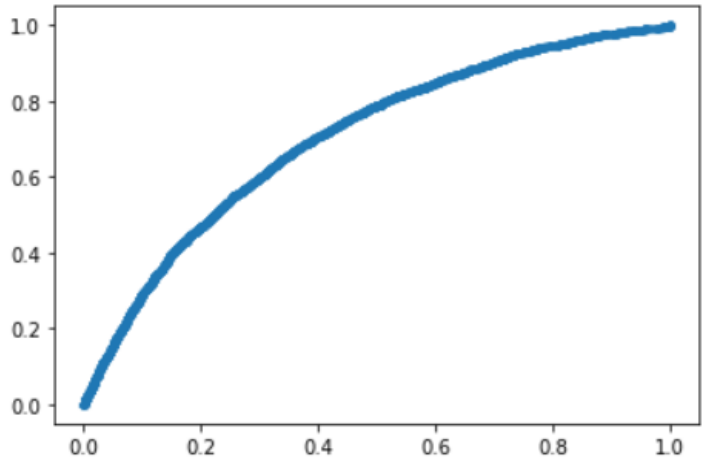
	Name of the Model	Score Accuracy
0	Random Forest	0.777656
1	Adaptive Boost Classifier	0.772332
2	KNN	0.750190
3	Naive Bayes: Gaussian	0.762106
4	Naive Bayes: Bernoulli	0.712752

```
(RandomForestClassifier(n_estimators=1000), 0.7062293844478784)
```

```
RFC Model : AUROC = 0.714
[<matplotlib.lines.Line2D at 0x13f11497610>]
```




```
ABC Model : AUROC = 0.705
[<matplotlib.lines.Line2D at 0x13f1155e280>]
```




```
(AdaBoostClassifier(n_estimators=300), 0.6933105844153683)
```

Optimization with hyperparameters and Gridsearch :

Comparaison of the ACC before Gridsearch and After:

	Name of the Model	Score ACC
0	Random Forest	0.777656
1	Adaptive Boost Classifier	0.772332
2	Random Forest Gridsearch	0.778078
3	Adaptive Boost Classifier Gridsearch	0.771487