Tuesday, May 03, 2022
Today's Paper

The New York Times

Georgetown University
Avi Arora
Sonali Pednekar
Vinayak Kannan
Olivier Kuhn De Chizelle

LIVE

Predicting Popularity of NYT Articles

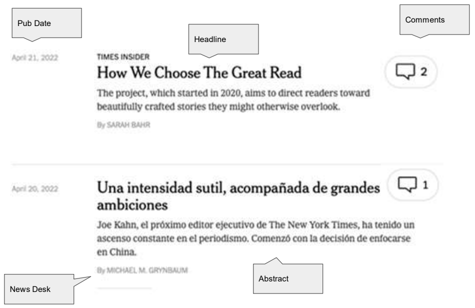
Data

The data utilized for this project was obtained on Kaggle and contains contains information about nearly 17,000 NYT articles published in 2020.

Objective

We want to predict whether a New York time article is popular or not.

We define popularity based on a threshold on the number of comments on the article



April 21, 2022

TIMES INSIDER

How We Choose The Great Read

The project, which started in 2020, aims to direct readers toward beautifully crafted stories they might otherwise overlook.

By SARAH BAHR

2

April 20, 2022

Una intensidad sutil, acompañada de grandes ambiciones

Joe Kahn, el próximo editor ejecutivo de The New York Times, ha tenido un ascenso constante en el periodismo. Comenzó con la decisión de enfocarse en China.

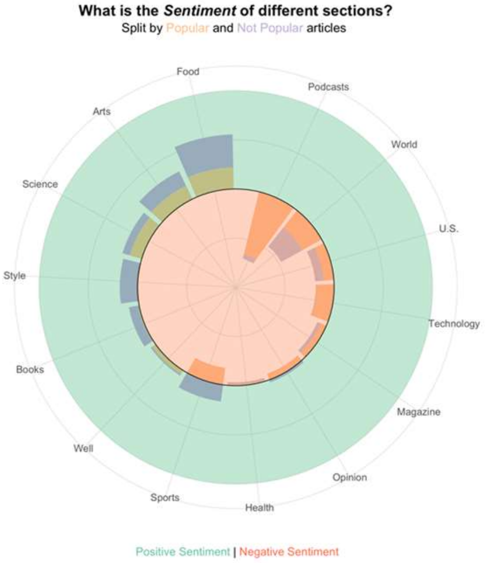
By MICHAEL M. GRYNBAUM

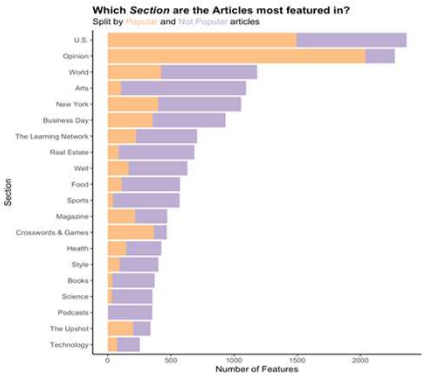
1

Feature Engineering

NLP

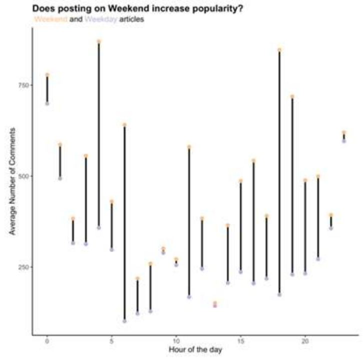
We utilized a sentiment intensity analyzer on headlines and abstracts in order to obtain an overall sentiment for each article.

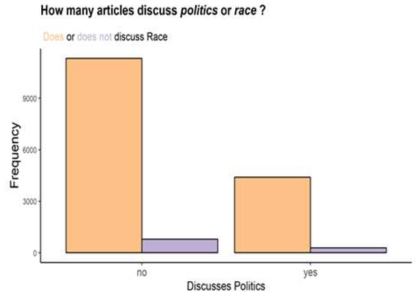

What is the *Sentiment* of different sections?
Split by Popular and Not Popular articles


Which *Section* are the Articles most featured in?
Split by Popular and Not Popular articles

Other Features

We generated flag variables to detect if an article discussed politics or racial issues.
We also broke down date and time of publication and created a flag for whether the article was published on a weekday or weekend.

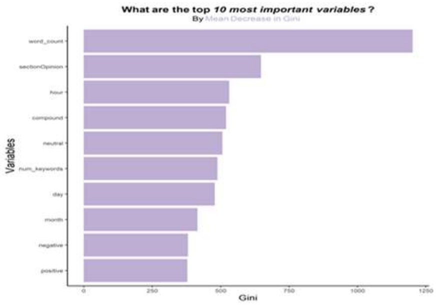

Does posting on Weekend increase popularity?
Weekend and Weekday articles


How many articles discuss *politics* or *race* ?
Does or does not discuss Race

Modelling

We created various models to best predict the accuracy of the NYT articles.
Among our methods are a Logistic Regression, LDA, Decision Tree, and Random Forest.

Metric	Logistic Regression	LDA	Decision Trees	Random Forest
Accuracy	76.70	77.54	78.10	79.20
Precision	78.00	66.00	78.00	80.00
Recall	76.00	78.00	76.00	79.00
F1 Score	77.00	71.00	76.00	73.00


What are the top *10 most important variables* ?
By Mean Decrease in Gini

Conclusion

Here are our recommendations to the NY times :

- The 'opinion' section seems to be the most popular section.
- Publishing an article between 11pm-2am happens to have the highest popularity.
- Articles in the food and arts section should have a positive sentiment to increase popularity
- Articles in the podcast section should negative to increase its popularity.