Predicting NYT Article Popularity

Benjamin Dornel - DSI 18

The New York Times

LIVE

House Delivers Impeachment Charge Against Trump to Senate

- The "incitement of insurrection" charge against former President Trump thrusts his political fate in the hands of 50 Republican senators.
- For now, they appear reluctant to convict him. Senators planned to wait until February to consider the case. Here's the latest in politics.

Nine House managers walked across the Capitol to inform the Senate that they were ready to prosecute former President Trump. Erin Schaff/The New York Times

Watchdog to Examine Whether Justice Dept. Helped Effort to Overturn Election

The inquiry was announced after revelations that former President Trump and another former official plotted to replace the acting attorney general.

Senate Confirms Yellen as Treasury Secretary as Stimulus Talks Loom

Janet Yellen, who won confirmation along bipartisan lines, now faces a big challenge in confronting a perilous economic threat.

A War Over Filibuster, a Stalling Tactic, Stops the Senate From the Start

Senator Mitch McConnell wants Democrats to promise not to gut the procedure that can grind the chamber to a halt.

Biden Sets in Motion Plan to Ban New Oil and Gas Drilling on Federal Land

President Biden is said to be planning several orders on climate change. A drilling ban would fulfill a campaign promise that infuriated the oil industry.

A pump jack and wind turbines in Stanton, Texas. Brandon Thibodeaux for The New York Times

Trump's Ban on Transgender Troops in Military Is Reversed

President Biden's action is part of his broader fight to prohibit discrimination against people based on their sexual orientation or gender identity.

President Biden increased his daily vaccination target to 1.5 million shots and extended virus-driven travel bans.

Opinion

Alexey Kovalev

Something Special Just Happened in Russia

Crackdown and coercion are no longer enough to stop people protesting.

Michelle Goldberg

Please, Biden, Try for 2 Million Shots a Day

The administration's vaccine plan isn't ambitious enough.

The New York Times

Opinion

I Want to Call the Capitol Rioters 'Terrorists.' Here's Why We Shouldn't.

New antiterrorism laws could end up targeting people of color.

By Adama Bah

Ms. Bah is an immigration activist.

Jan. 25, 2021

438

Comments 438

The comments section is closed. To submit a letter to the editor for publication, write to letters@nytimes.com.

NYT Picks | Reader Picks | All

Kyle C — Times Pick

Chicago | 7h ago

If the government wants to successfully prosecute and minimize the damage of domestic terrorist organizations, it doesn't need a new law that further infringes on our civil liberties, it just needs more Twitter accounts. The Jan. 6 attack was wide out in the open. It was on Facebook, Twitter, and I'm assuming other social media sites like Parler that I don't bother to look at.

2 Replies 31 Recommend Share Flag

Patrick — Times Pick

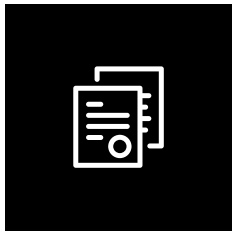
NYC | 8h ago

All the war on terror achieved was to strip away more liberties from law abiding citizens. It was a failure. Yet we all line up at airports dutifully getting scanned with no proof it will stop another attack and that's the least of it. The Patriot Act should be changed to The Fascist Act. Interestingly enough, all we have achieved is to become more like the enemy and the repressive societies we are fighting against.

7 Replies 92 Recommend Share Flag

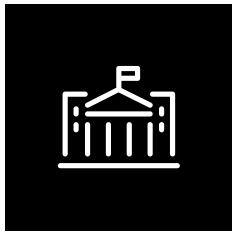
Enter the New York Times

Pivoting from print & achieving digital success



High Quality Content

Meeting the demand for high quality, original, independent journalism



Community Engagement

Building reader loyalty through engagement

Embracing Technology

Hiring journalists who can code & creating interactive features

The New York Times

Online articles from January — December 2020

Total Articles

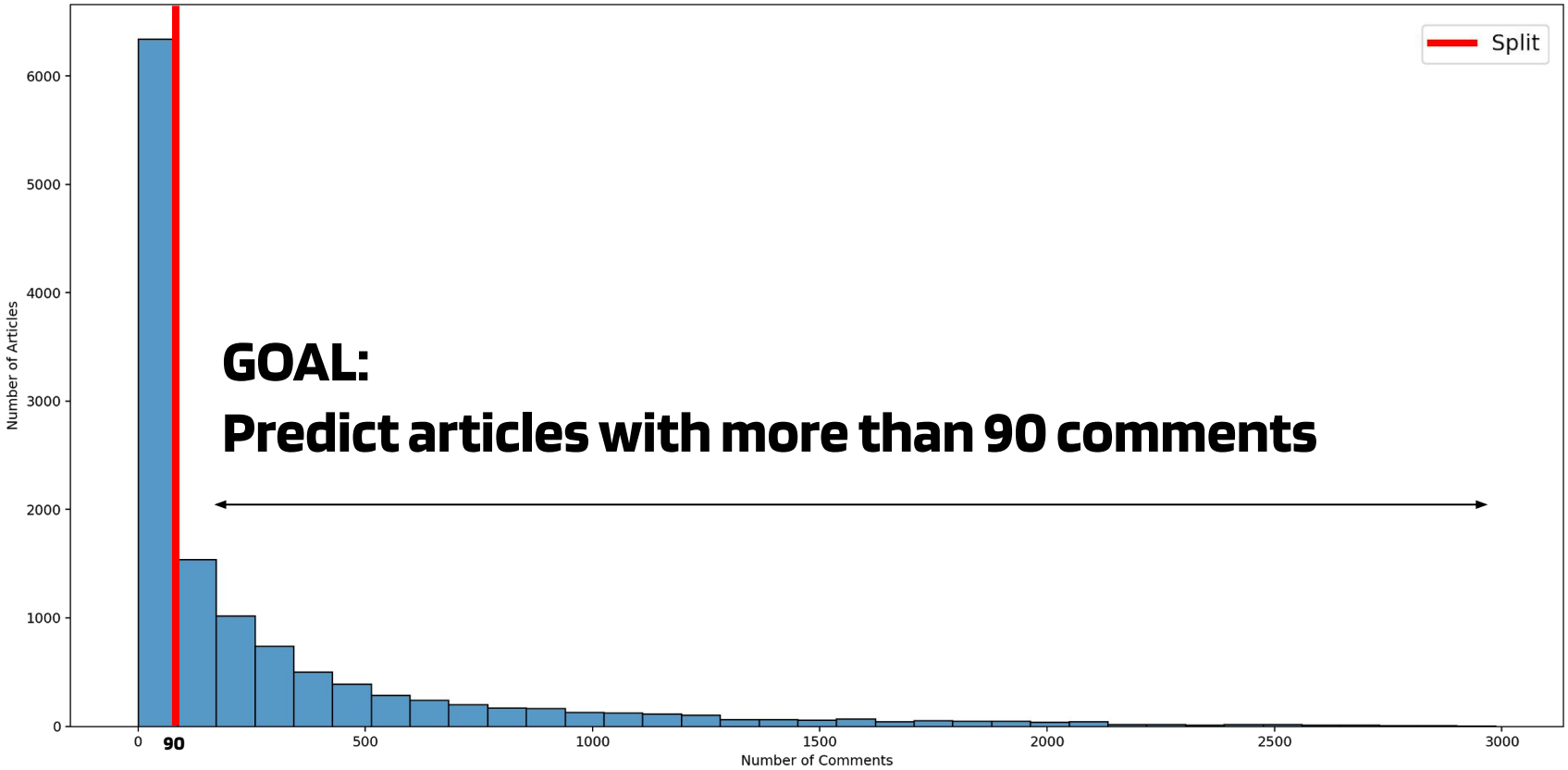
16K

Total Comments

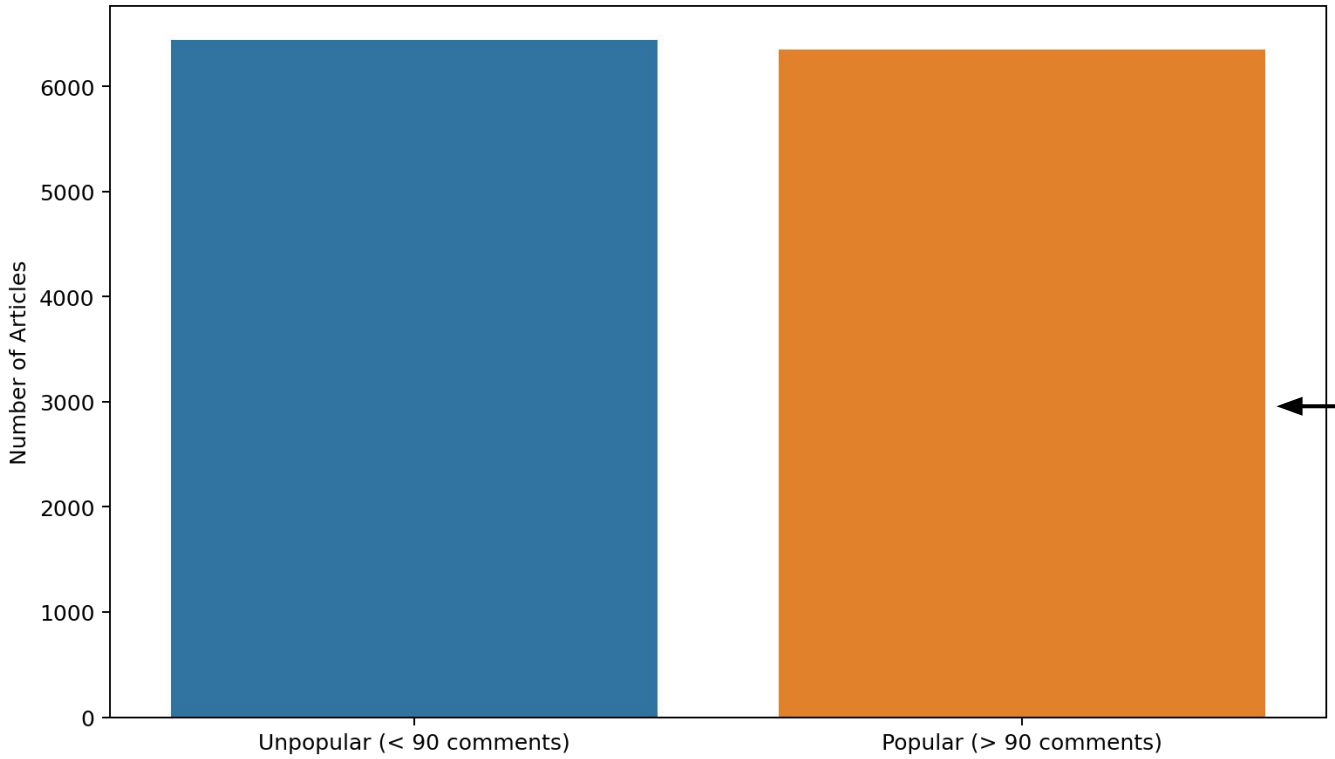
5M

Let's look at how comments are spread across articles

+ Comments Distribution



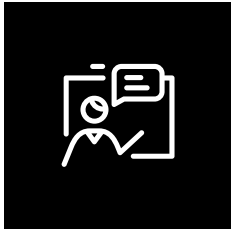
+ Binary Classification



GOAL: predict positive class

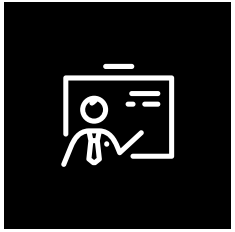
+ Objectives

What do we want to accomplish?



Business Objective

How can the NYT staff improve article popularity?



Data Science Objective

Can we accurately predict popularity? What are the most important predictors?

+ Data Overview

What are we working with?

News Desk	Primary category for articles. Each desk has a unique focus. 60 unique.
Section	Secondary category. Further separates news into different groups. 41 unique.
Subsection	Tertiary category. Tends to cover more niche areas. 61 unique.
Material	Type of material. e.g. obituary, review, editorial, news analysis etc.
Keywords	People / places / organizations / things listed within the article

+ Data Overview

What are we working with?

Headline	Article headline
Abstract	Article abstract -- contains a rough summary of the article
Word Count	Number of words in article
Publication Date	Timestamp date that article was published

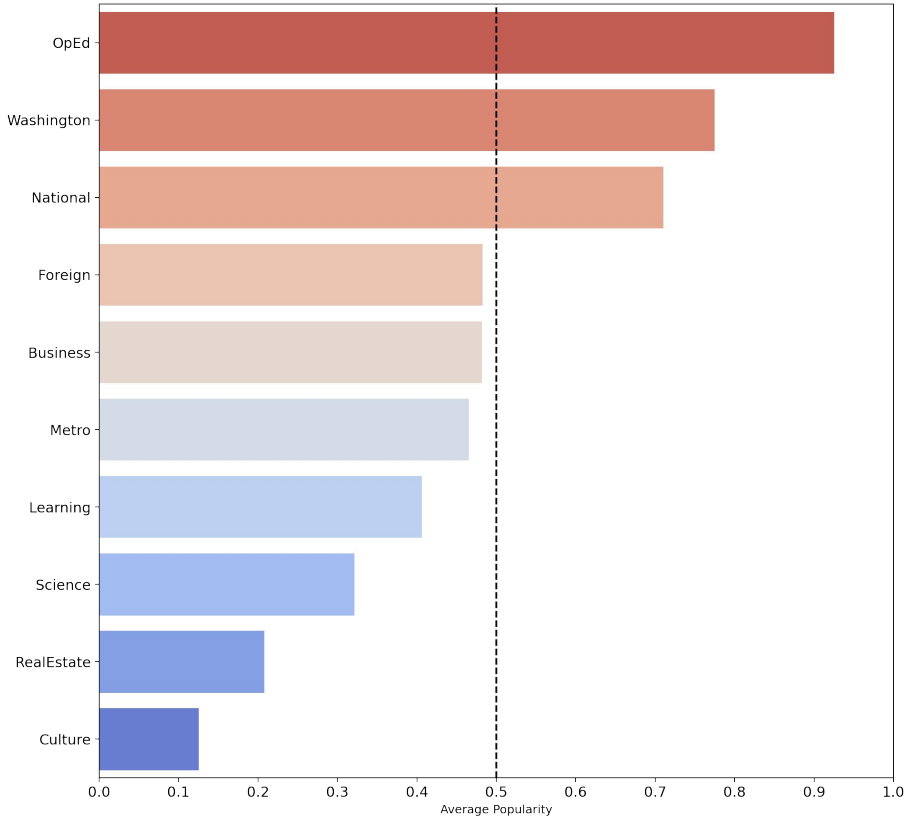
How do we predict an article's performance?

+ EDA - News Desk Popularity

News desks are a strong predictor of article popularity.

- These news desks contain over half of all articles in the datasets.
- Articles published in OpEd news desk tend to have a higher number of comments.
- Local U.S. news tends to be more popular than foreign news

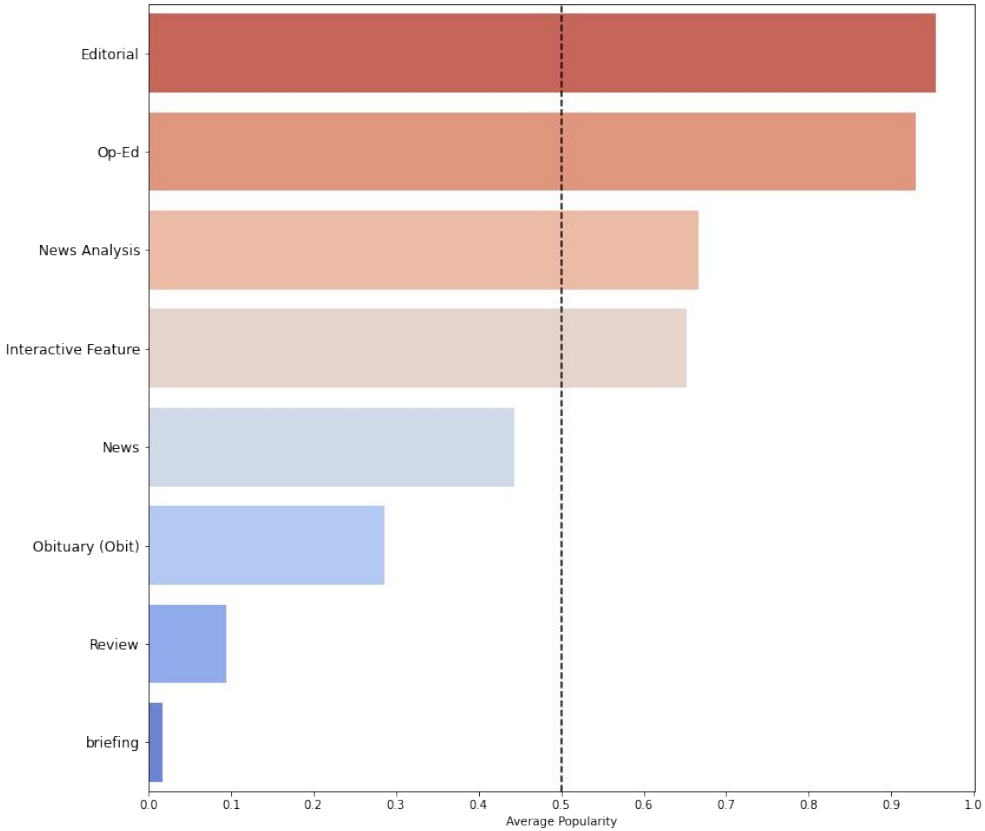
We can also look at section / subsection as additional ways to gauge popularity.



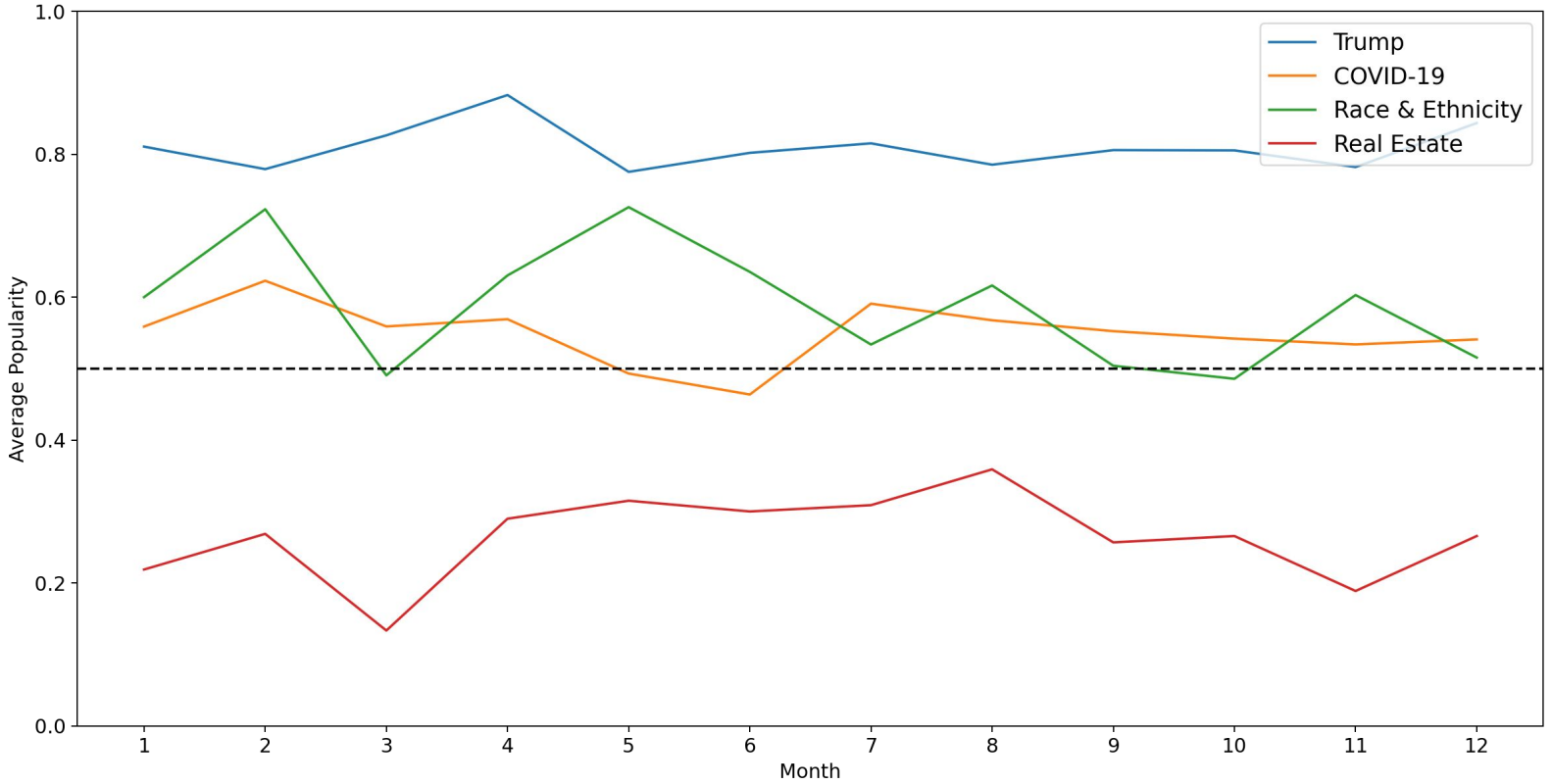
+ EDA - Material Popularity

The type of material also affects an article's popularity.

- Editorials and Opinion pieces are the most popular
- News analysis and interactive features are moderately popular
- Regular news, obituaries and reviews tend to be more unpopular.

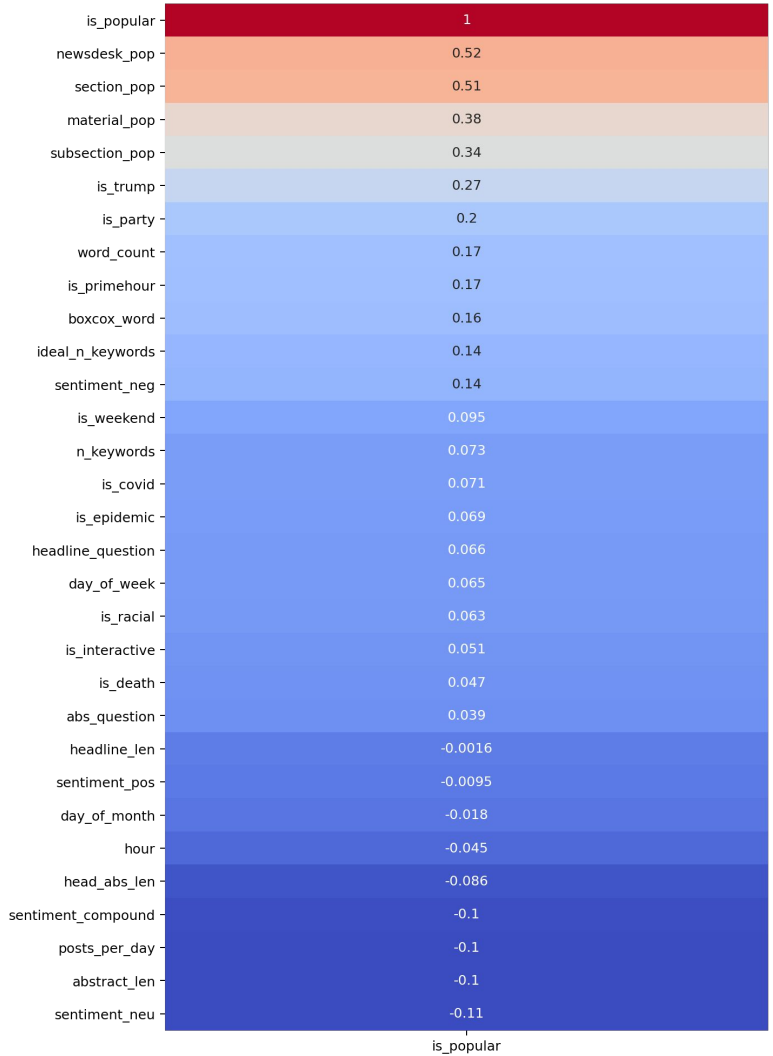


+ EDA – Keyword Popularity



+ Feature Engineering

- Box-cox transformation of word count
- Time variables based on publication date
  - Day of week, day of month, hour, weekend
  - Number of articles posted per day
- Keywords
  - Donald Trump
  - COVID-19
- Headline / abstract length
- Whether headline / abstract contains a question mark
- Convert categorical features to ordinal features based on average popularity
- Sentiment based on headline + abstract
- Topic clustering based on headline



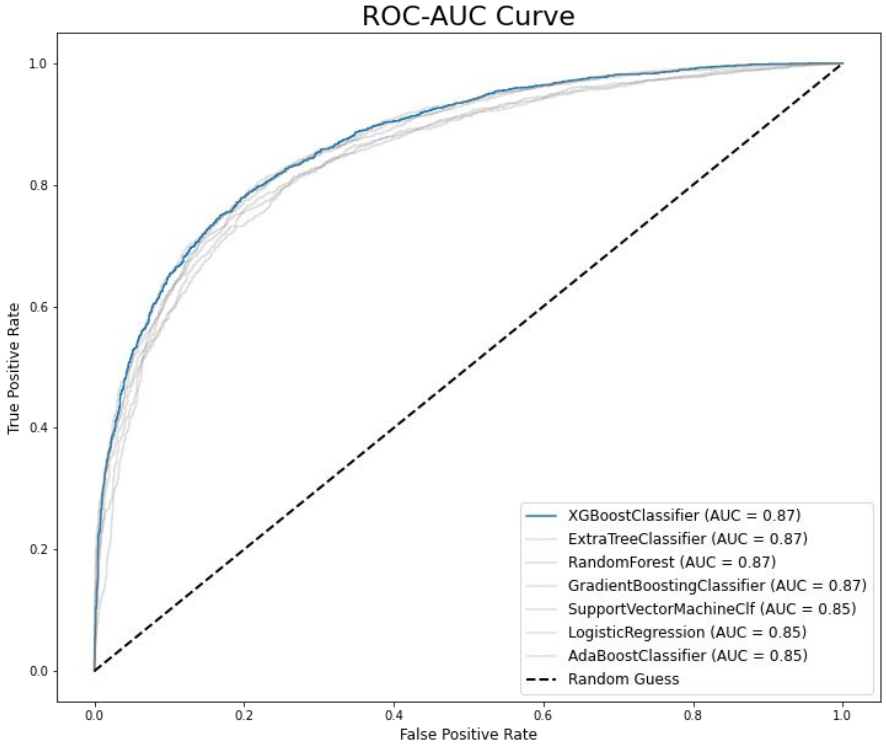
+ Modelling

- Best model:

  XGBoost Classifier

  - Accuracy: 0.78
  - Precision: 0.77
  - Recall: 0.77
  - ROC-AUC: 0.87

- Large increase over baseline model accuracy of 0.49

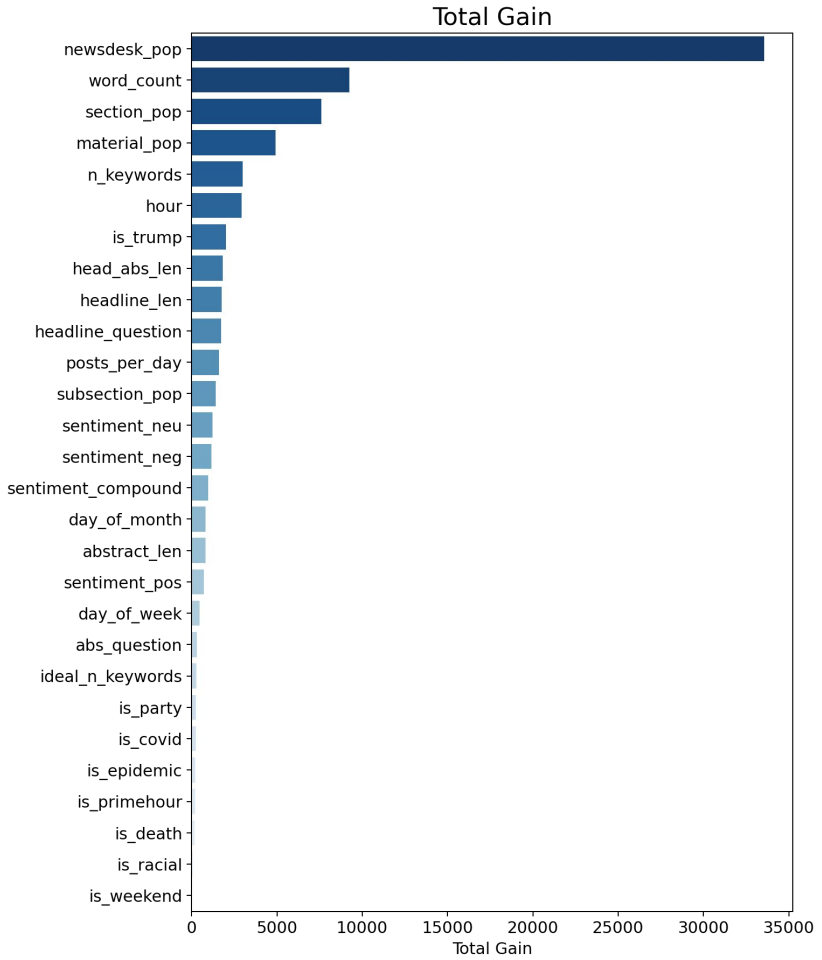


So what does our model tell us?

+ Model Insights

Most important features:

1. News desk
2. Word count (regular)
3. Section
4. Material type
5. Number of keywords
6. Hour
7. Subsection
8. Headline length
9. Combined headline & abstract length
10. Sentiment
11. Keyword: Donald J. Trump



+ Recommendations

- Using the model, we can now predict which articles are more or less likely to be popular. This will **increase moderation efficiency** and increase comment approval speed.

- For articles that our model predicts to be unpopular:
  - Re-write 'neutral' headlines & abstracts
  - Shorten length of headlines & abstracts and include a question mark
  - Change publication time to between 10pm and 3am
  - Tie the story to the current socio-political context and use the right keywords
  - Use a recommendation system to drive traffic to from popular articles to unpopular articles

+ Limitations

- Just because an article has a low number of comments doesn't mean it's a low-performing article. The article might not appeal to regular commentators, and might do better on social media.

- The NYT has to also uphold a degree of journalistic integrity -- even though using a clickbait headline might get them more comments, it's in their own interest to remain an impartial and trusted source of news.

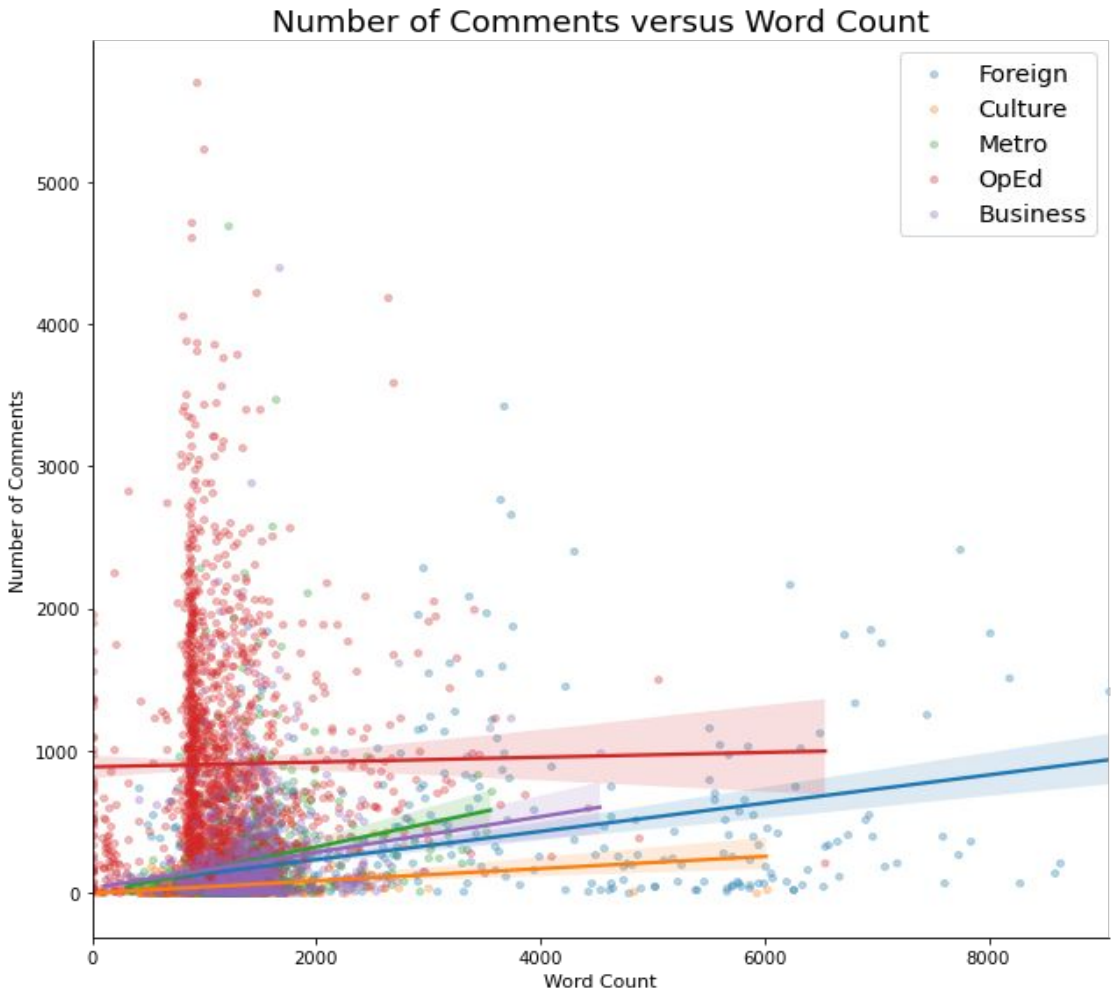
- This model is heavily affected by an article's news desk / section / subsection. Topic 'freshness' might be a strong factor that our model overlooks.

End

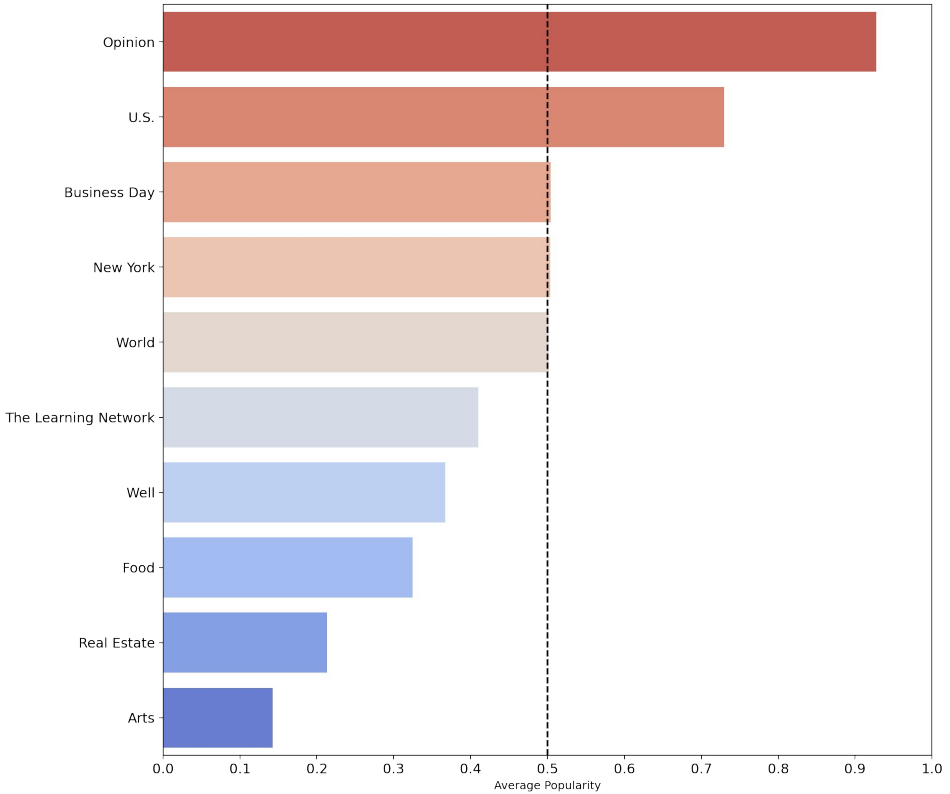
Appendix A

Additional Visualizations

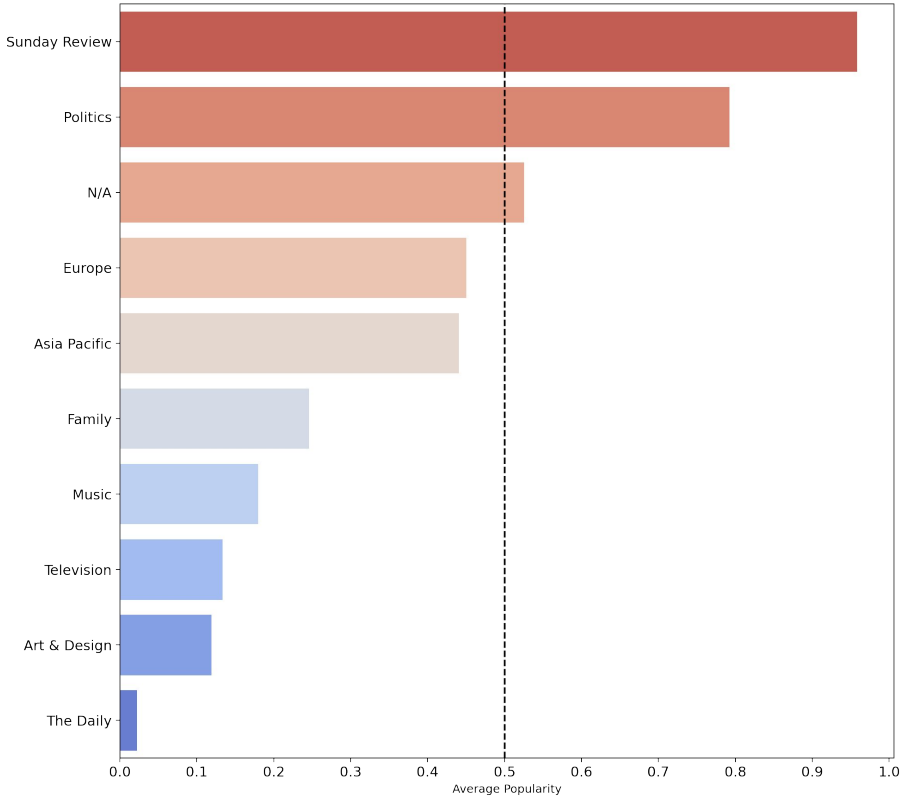
- Generally positive trend except for articles from OpEd newsdesk



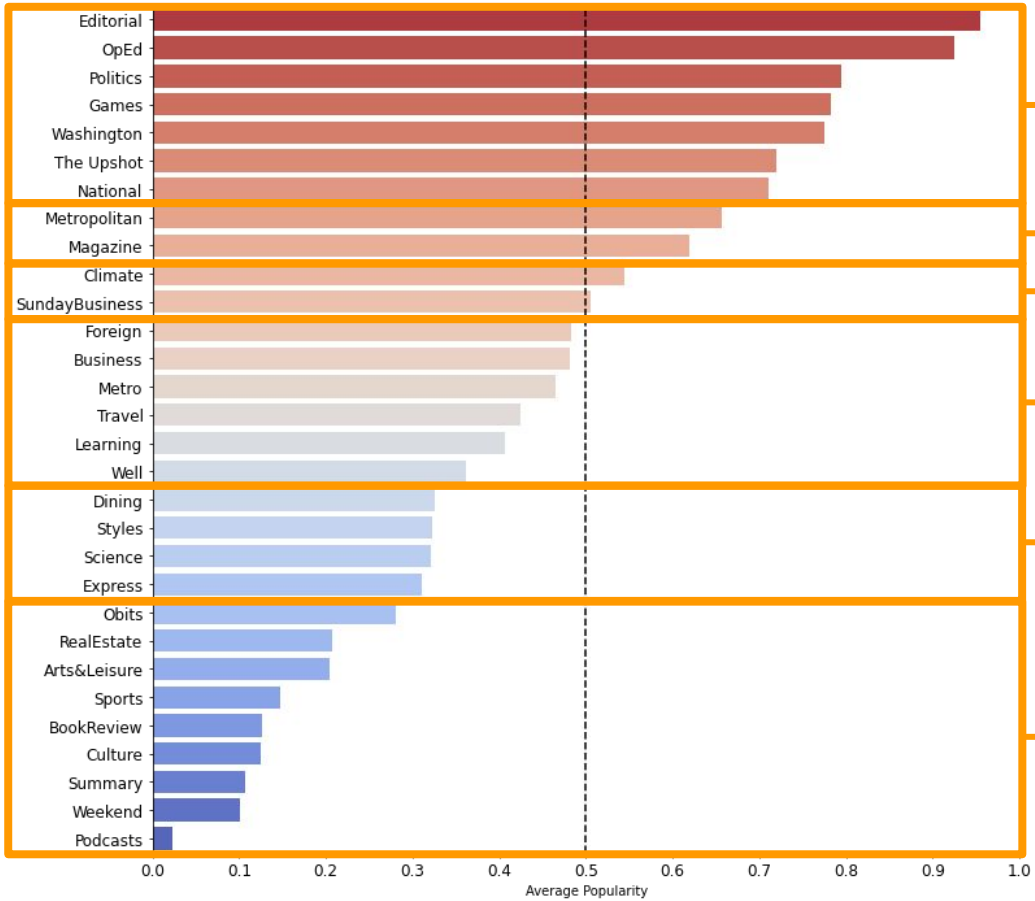
+ Section Popularity



+ Subsection Popularity



Categorical to Ordinal Mapping



Average Pop. >= 0.7 → **5**

0.7 < Average Pop. >= 0.6 → **4**

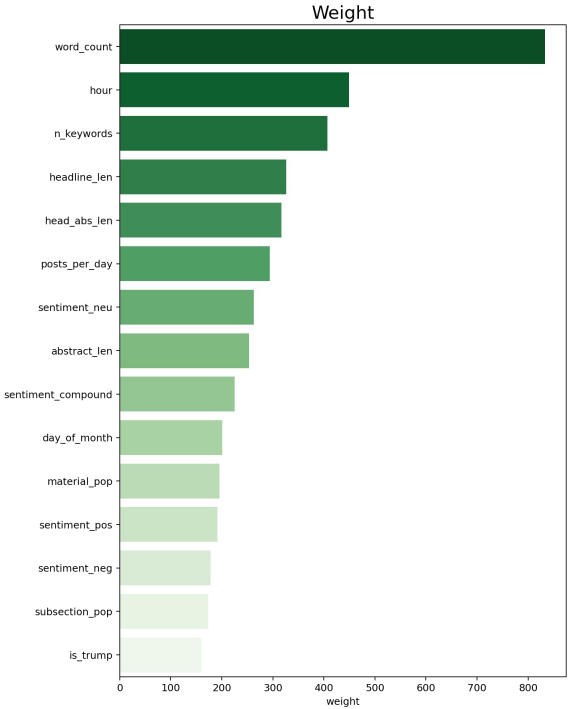
0.6 < Average Pop. >= 0.5 → **3**

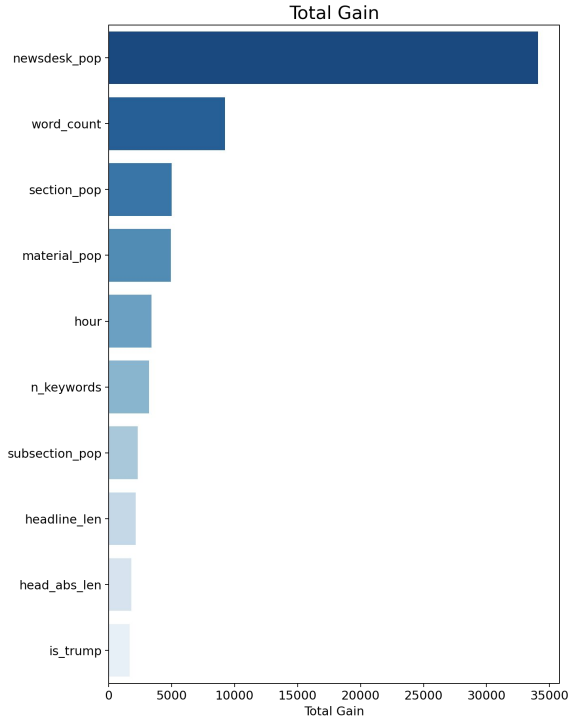
0.5 < Average Pop. >= 0.4 → **2**

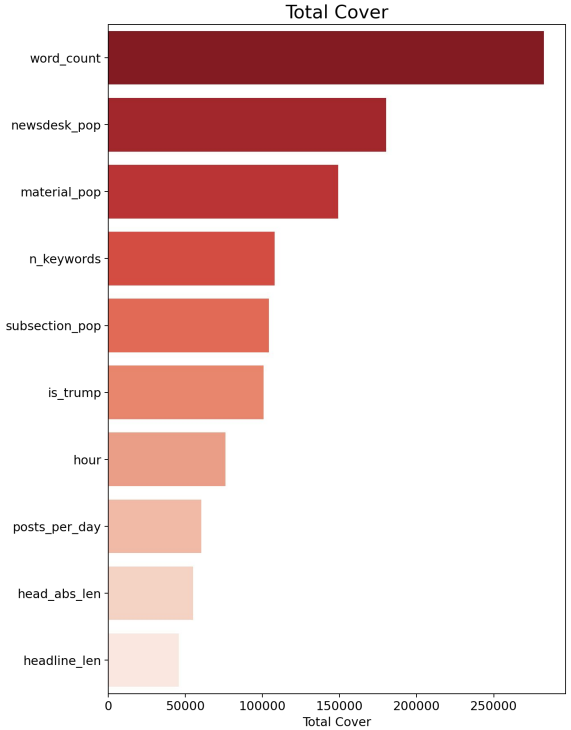
0.4 < Average Pop. >= 0.3 → **1**

Average Pop. < 0.3 → **-1**

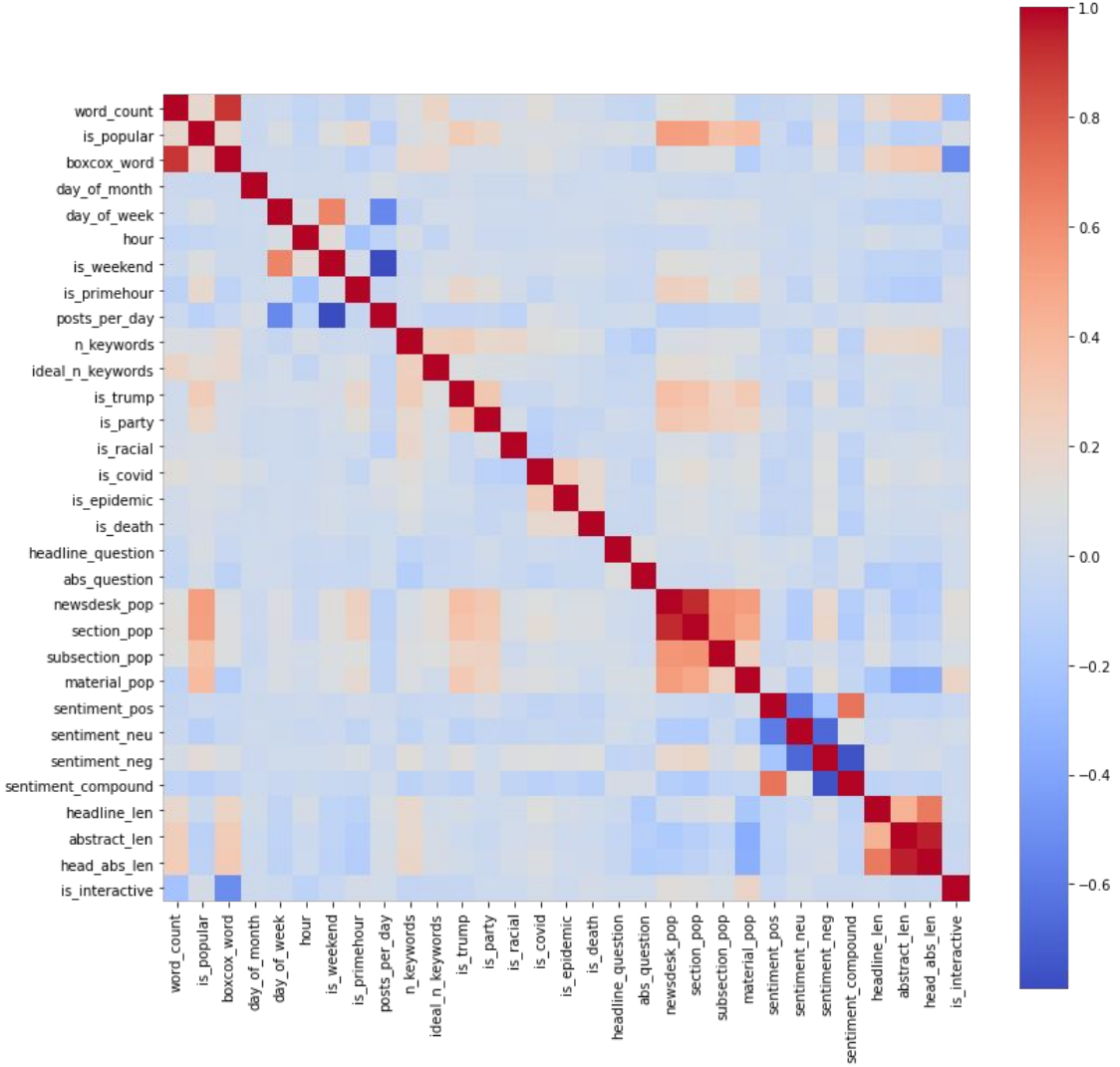
+ Model Insights







Feature Heatmap



Appendix B

Additional Resources

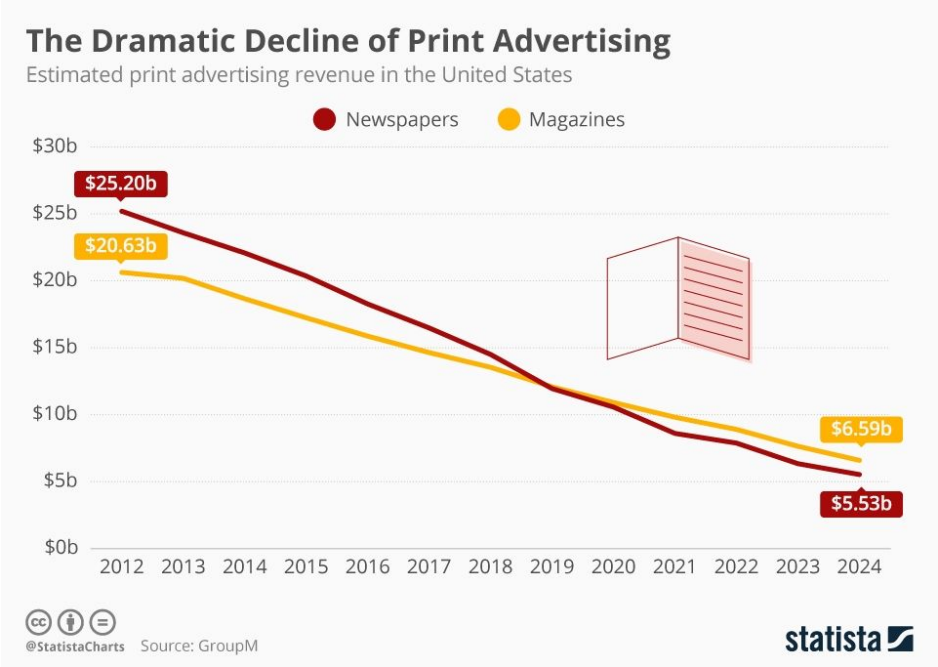
+ Background

Why does online news matter?

The newspaper industry isn't doing well:

- Huge decline in advertising $
- Falling print subscriptions
- Continuous layoffs since 2008

Massive business strategy changes & overhauls are required to stay profitable.



+ Misclassifications (False Positive)

	actual	predicted	proba	newsdesk	section	subsection	headline	abstract	keywords
254	0.0	1.0	0.937960	OpEd	Opinion	NaN	What We Can Learn From the Rise and Fall of 'Political Blackness'	Before "people of color" and BIPOC, Britain had its own inclusive concept for nonwhite people.	[Black People, Blacks, Race and Ethnicity, Minorities, Great Britain, Politics and Government]
66	0.0	1.0	0.931563	Politics	U.S.	Politics	Democrats Mount an All-Out Effort to Get Detroit to Vote	Urgently trying to avoid a repeat of 2016, when Donald Trump won Michigan by a razor-thin margin, Democrats have bee...	[Presidential Election of 2020, Black People, Blacks, Biden, Joseph R Jr, Harris, Kamala D, Trump, Donald J, Detroit...
321	0.0	1.0	0.844512	Washington	U.S.	Politics	Under Pence, Politics Regularly Seeped Into the Coronavirus Task Force	In taking a leading role in managing the White House's response to the pandemic, the vice president and his team had...	[Pence, Mike, Coronavirus (2019-nCoV), United States Politics and Government, Presidential Election of 2020, Trump, ...
190	0.0	1.0	0.837098	The Upshot	The Upshot	NaN	The Pandemic Has Hindered Many of the Best Ideas for Reducing Violence	That may be one part of this year's rise in violent crime.	[Crime and Criminals, Therapy and Rehabilitation, Nonprofit Organizations, Shutdowns (Institutional), Coronavirus (2...
318	0.0	1.0	0.829171	Metropolitan	New York	NaN	Their Buzzy Off Broadway Play Shut Down. Here's What They Did Next.	When the virus forced New York theater to go dark, it upended thousands of lives, from actors to ticket takers. An a...	[Theater, New York City, Playwrights Horizons, Greenfield, Adam, Khoury, Sylvia, Selling Kabul (Play), Quarantine (L...

+ Misclassifications (False Negative)

	actual	predicted	proba	newsdesk	section	subsection	headline	abstract	keywords
34	1.0	0.0	0.094006	Weekend	Arts	Music	The Special Place Where Ella Fitzgerald Comes Alive	The singer's concert recordings have always had a power that her studio outings could only imply. "Ella: The Lost Be...	[Jazz, Fitzgerald, Ella, Ella: The Lost Berlin Tapes (Album)]
50	1.0	0.0	0.105682	RealEstate	Real Estate	NaN	Real Estate Sales Continue to Stagnate in Manhattan	The number of apartments sold in the last three months was down by 46 percent compared to the same period in 2019.	[Real Estate and Housing (Residential), Quarantine (Life and Culture), Shutdowns (Institutional), Condominiums, Manh...
33	1.0	0.0	0.237722	Well	Well	Live	Where You Carry Body Fat May Affect How Long You Live	Extra weight in some places may lower your risk of dying prematurely.	[Weight, Deaths (Fatalities)]
60	1.0	0.0	0.247927	Styles	Style	NaN	'Ridicule': The French Reaction to 'Emily in Paris'	Darren Star's latest serial goes down like sour wine for actual Parisians.	[Television, Emily in Paris (TV Program), Paris (France), your-feed-fashion]
17	1.0	0.0	0.298873	Well	Well	Mind	Laughter May Be Effective Medicine for These Trying Times	Doctors, nurses and therapists have a prescription for helping all of us to get through these difficult times: Try a...	[Content Type: Service, Anxiety and Stress, Laughter, Comedy and Humor, Emergency Medical Treatment, Hospitals, Coro...