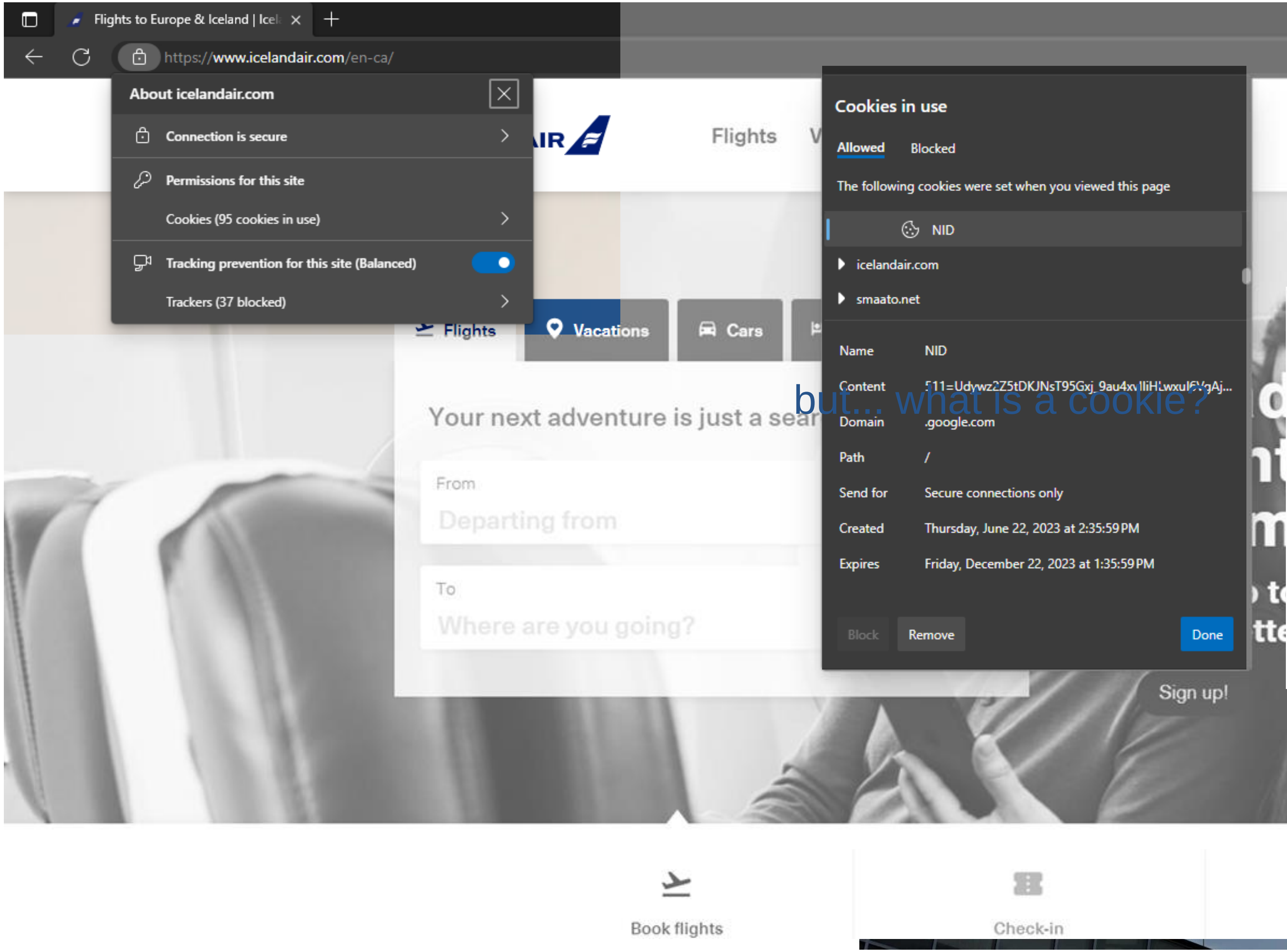
Contextual Ads

as the New Frontier in Privacy Compliant Advertising

Contents

- Current digital advertising landscape - what’s at stake?
- How can Machine Learning in contextual advertising contribute to addressing privacy concerns?
- What is the impact in terms of dollars ?
- What’s our data - how does it look like?
- Pre-processing
- Baseline Model
- Next steps…

Current digital advertising landscape - what's at stake?



Flights to Europe & Iceland | Icel

https://www.icelandair.com/en-ca/

About icelandair.com

- Connection is secure
- Permissions for this site
- Cookies (95 cookies in use)
- Tracking prevention for this site (Balanced)
- Trackers (37 blocked)

Cookies in use

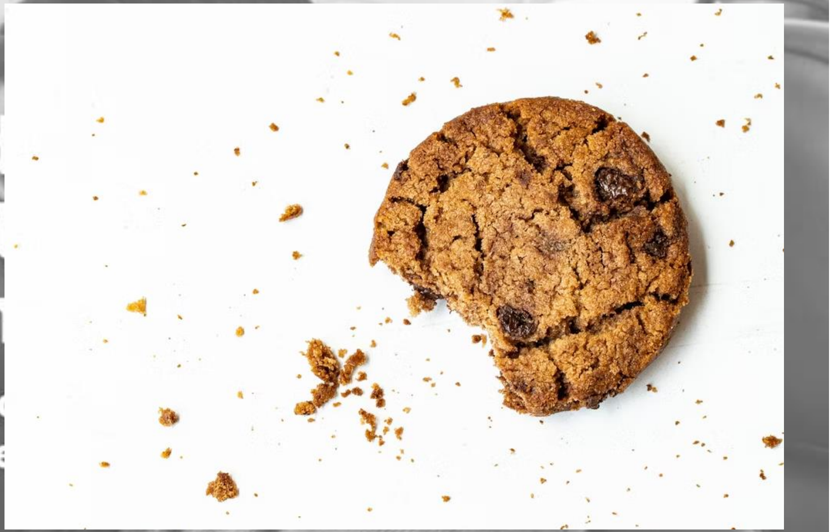
Allowed | Blocked

The following cookies were set when you viewed this page

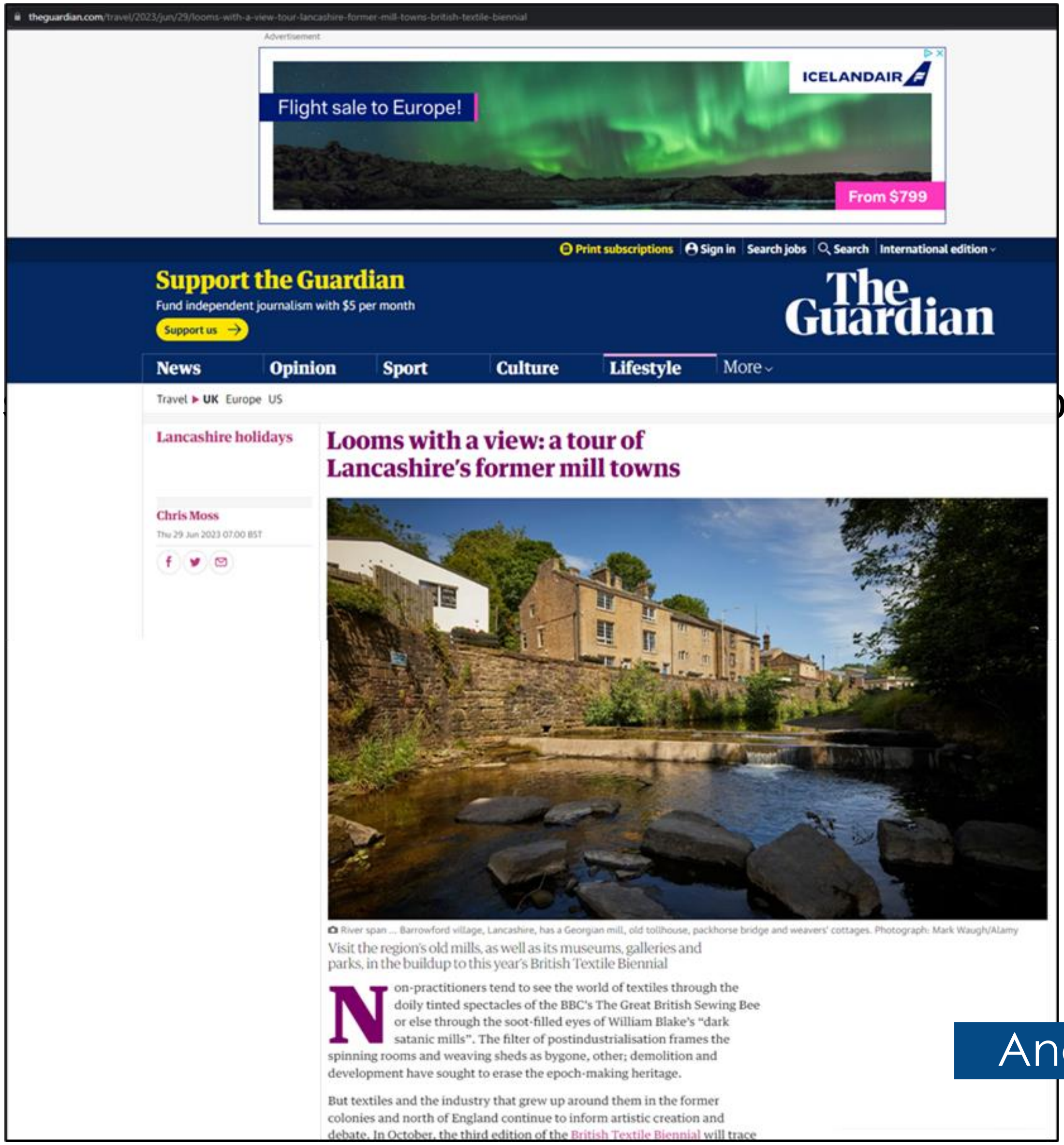
- NID
- icelandair.com
- smaato.net

Name	NID
Content	511=Udywz2Z5tDKJNsT95Gxj_9au4xvlliHLwxul6VgAj...
Domain	.google.com
Path	/
Send for	Secure connections only
Created	Thursday, June 22, 2023 at 2:35:59 PM
Expires	Friday, December 22, 2023 at 1:35:59 PM

but... what is a cookie?



How can Machine Learning in contextual advertising contribute to addressing privacy concerns?



theguardian.com/travel/2023/jun/29/looms-with-a-view-tour-lancashire-former-mill-towns-british-textile-biennial

Advertisement

ICELANDAIR

Flight sale to Europe!

From $799

Print subscriptions | Sign in | Search jobs | Search | International edition

Support the Guardian

Fund independent journalism with $5 per month

Support us →

The Guardian

News | Opinion | Sport | Culture | Lifestyle | More

Travel ▶ UK Europe US

Lancashire holidays

Looms with a view: a tour of Lancashire's former mill towns

Chris Moss

Thu 29 Jun 2023 07.00 BST

River span … Barrowford village, Lancashire, has a Georgian mill, old tollhouse, packhorse bridge and weavers' cottages. Photograph: Mark Waugh/Alamy

Visit the region's old mills, as well as its museums, galleries and parks, in the buildup to this year's British Textile Biennial

Non-practitioners tend to see the world of textiles through the doily tinted spectacles of the BBC's The Great British Sewing Bee or else through the soot-filled eyes of William Blake's "dark satanic mills". The filter of postindustrialisation frames the spinning rooms and weaving sheds as bygone, other; demolition and development have sought to erase the epoch-making heritage.

But textiles and the industry that grew up around them in the former colonies and north of England continue to inform artistic creation and debate. In October, the third edition of the British Textile Biennial will trace

Propo f a cookie (tracking users across w b) to show an ad?

Travel Ad

Picks advertisers (IcelandAir) who wants to show travel ad

Ad platform receives the category signal Example: Travel

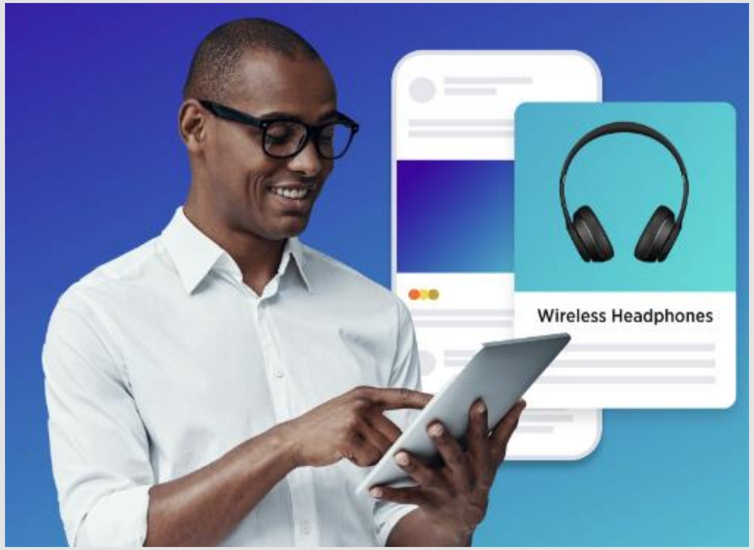
Analyze texts

ML Predicts the category it belongs

What is the impact in terms of dollars ?

“..the Interactive Advertising Bureau estimates publishers' exposure at up to **$10 billion in lost revenue** without third party **cookies**..” Source: IAB

Impact of our proposed solution



Users seeing relevant ads

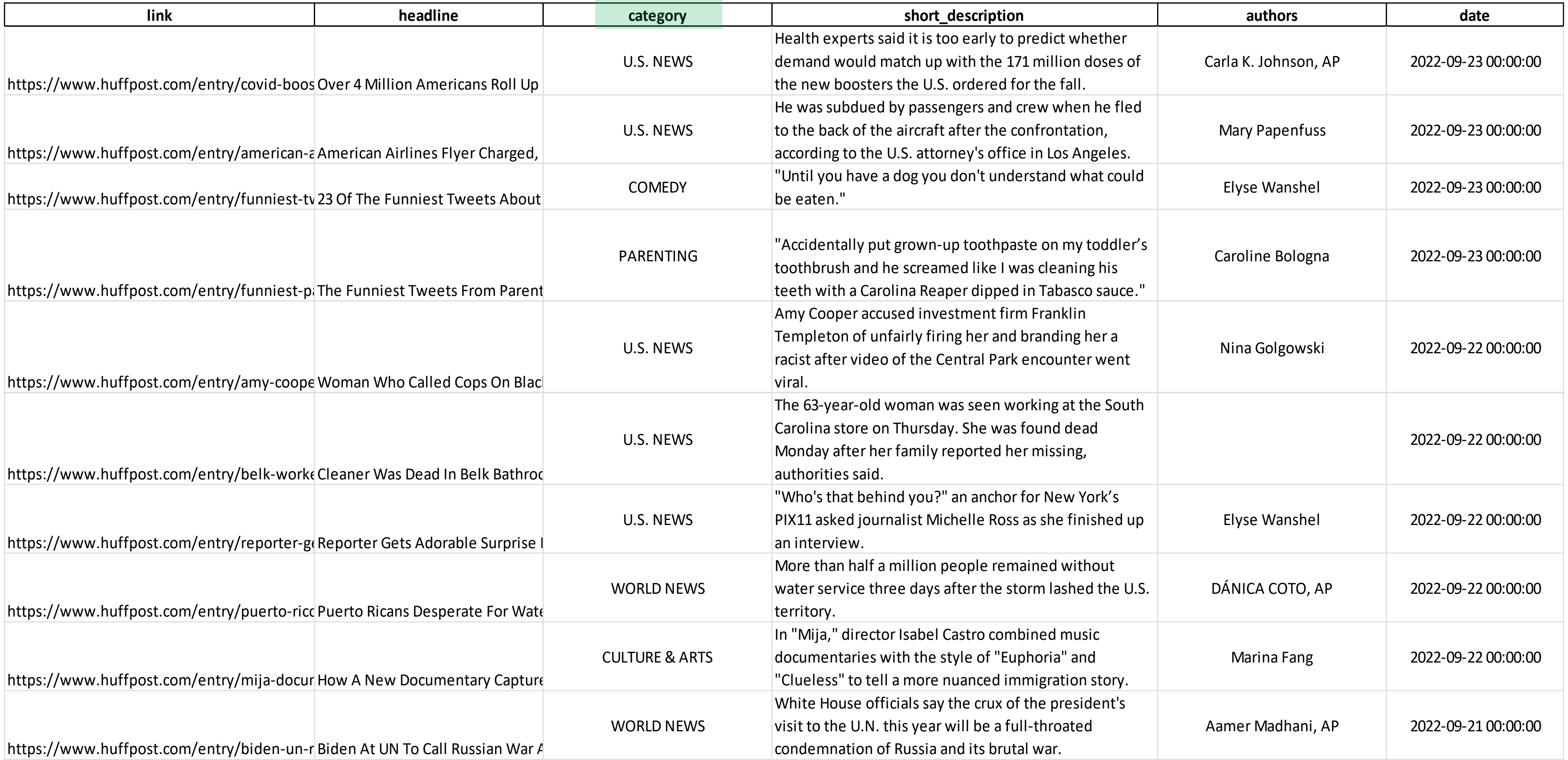
Advertisers will gain high ROI

Publishers sell more ads

What’s our data - how does it look like?

This dataset contains news headlines from 2012 to 2022 from HuffPost.

209,527 rows
&
6 columns

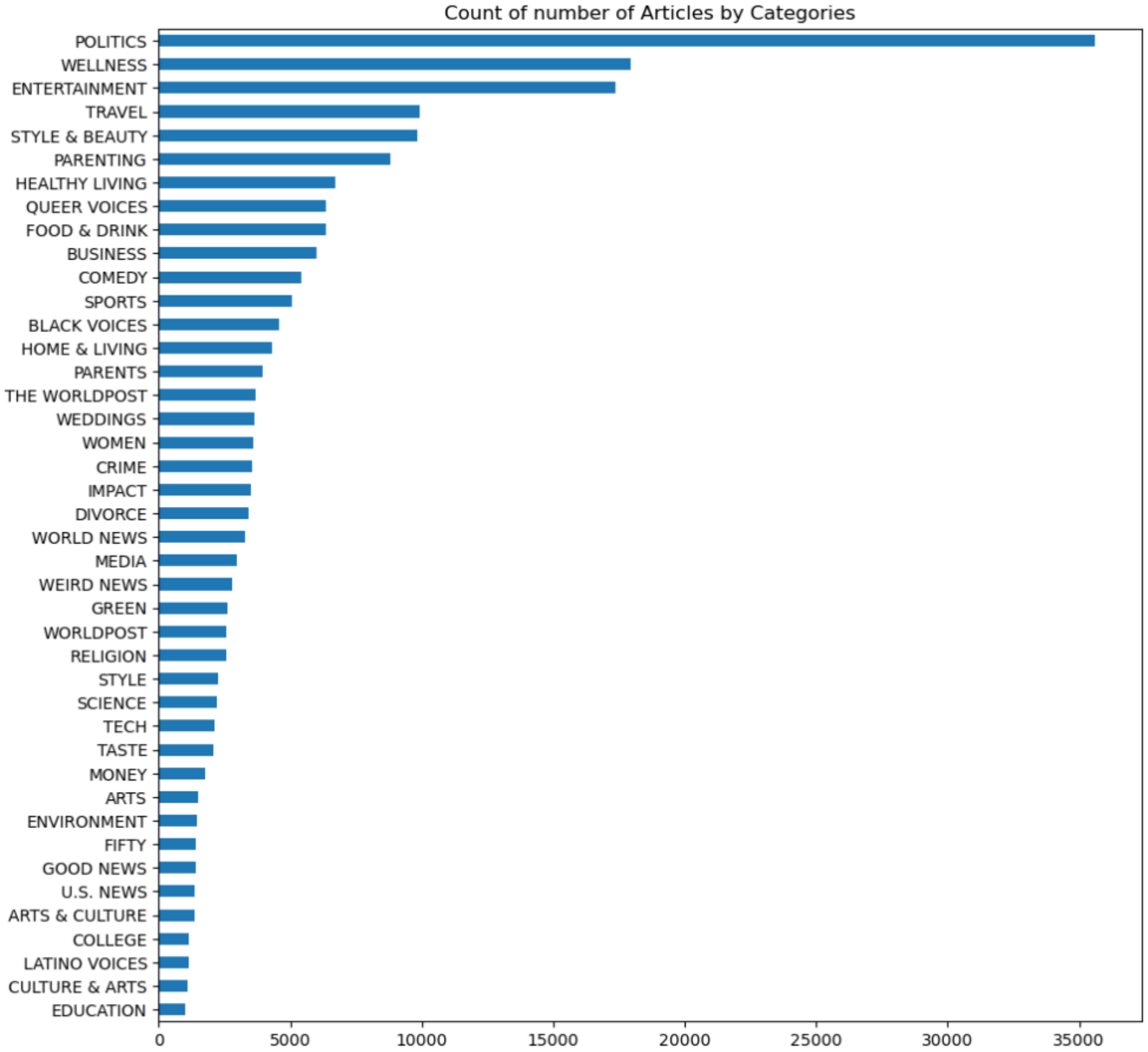


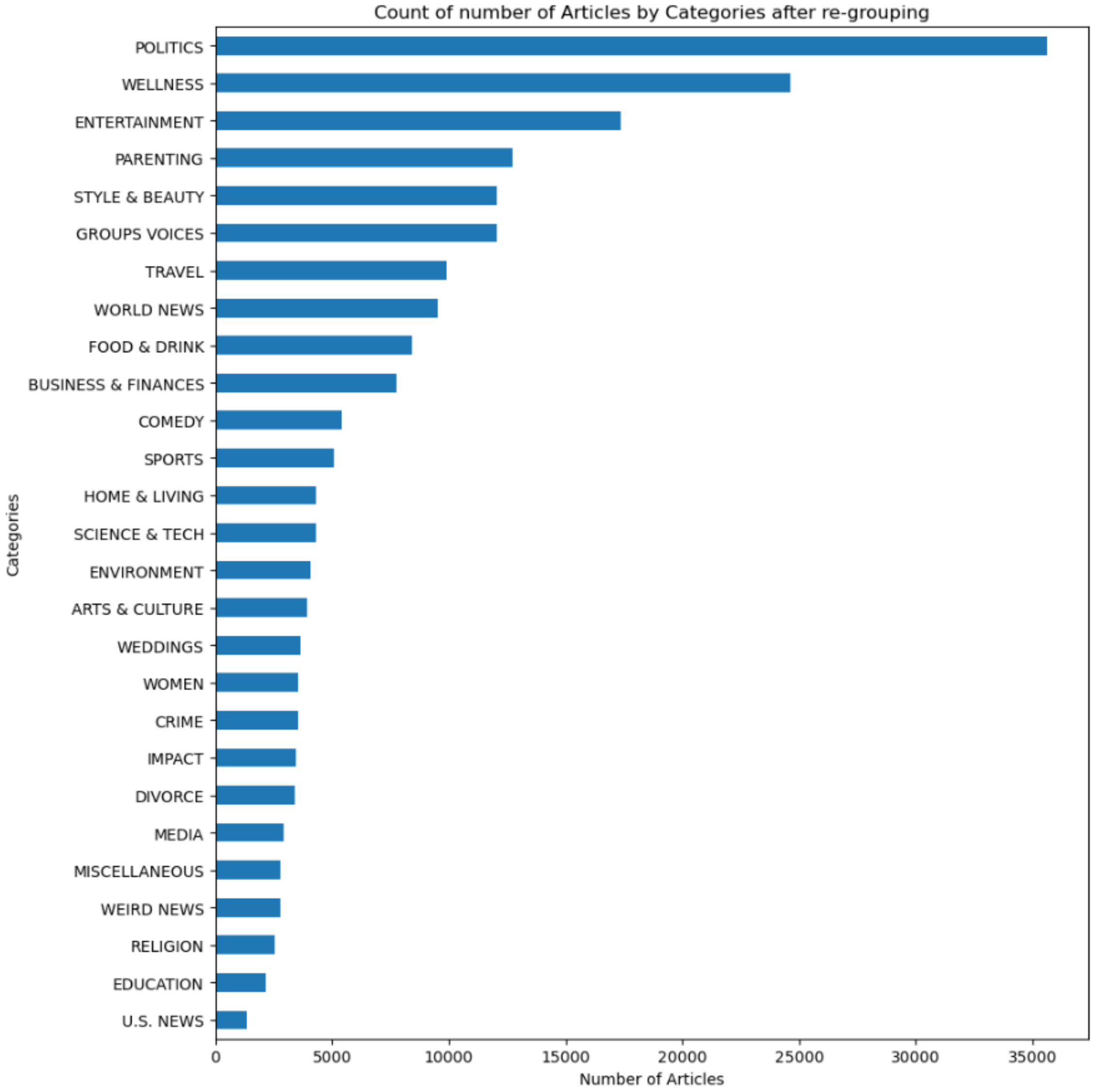
link	headline	category	short_description	authors	date
https://www.huffpost.com/entry/covid-boos	Over 4 Million Americans Roll Up	U.S. NEWS	Health experts said it is too early to predict whether demand would match up with the 171 million doses of the new boosters the U.S. ordered for the fall.	Carla K. Johnson, AP	2022-09-23 00:00:00
https://www.huffpost.com/entry/american-a	American Airlines Flyer Charged,	U.S. NEWS	He was subdued by passengers and crew when he fled to the back of the aircraft after the confrontation, according to the U.S. attorney's office in Los Angeles.	Mary Papenfuss	2022-09-23 00:00:00
https://www.huffpost.com/entry/funniest-tv	23 Of The Funniest Tweets About	COMEDY	"Until you have a dog you don't understand what could be eaten."	Elyse Wanshel	2022-09-23 00:00:00
https://www.huffpost.com/entry/funniest-pa	The Funniest Tweets From Parent	PARENTING	"Accidentally put grown-up toothpaste on my toddler’s toothbrush and he screamed like I was cleaning his teeth with a Carolina Reaper dipped in Tabasco sauce."	Caroline Bologna	2022-09-23 00:00:00
https://www.huffpost.com/entry/amy-coope	Woman Who Called Cops On Blacl	U.S. NEWS	Amy Cooper accused investment firm Franklin Templeton of unfairly firing her and branding her a racist after video of the Central Park encounter went viral.	Nina Golgowski	2022-09-22 00:00:00
https://www.huffpost.com/entry/belk-worke	Cleaner Was Dead In Belk Bathroc	U.S. NEWS	The 63-year-old woman was seen working at the South Carolina store on Thursday. She was found dead Monday after her family reported her missing, authorities said.		2022-09-22 00:00:00
https://www.huffpost.com/entry/reporter-ge	Reporter Gets Adorable Surprise I	U.S. NEWS	"Who's that behind you?" an anchor for New York’s PIX11 asked journalist Michelle Ross as she finished up an interview.	Elyse Wanshel	2022-09-22 00:00:00
https://www.huffpost.com/entry/puerto-ricc	Puerto Ricans Desperate For Wate	WORLD NEWS	More than half a million people remained without water service three days after the storm lashed the U.S. territory.	DÁNICA COTO, AP	2022-09-22 00:00:00
https://www.huffpost.com/entry/mija-docur	How A New Documentary Capture	CULTURE & ARTS	In "Mija," director Isabel Castro combined music documentaries with the style of "Euphoria" and "Clueless" to tell a more nuanced immigration story.	Marina Fang	2022-09-22 00:00:00
https://www.huffpost.com/entry/biden-un-r	Biden At UN To Call Russian War A	WORLD NEWS	White House officials say the crux of the president's visit to the U.N. this year will be a full-throated condemnation of Russia and its brutal war.	Aamer Madhani, AP	2022-09-21 00:00:00

Feature	% missing
link	0%
headline	0%
category	0%
short_description	9%
authors	18%
date	0%

Pre-processing

Effort to reduce class imbalance: From 42 → 27 unique categories





Pre-processing

Custom-Tokenizer:

- remove punctuation and set to lower case
- remove digits using list comprehension
- remove html tags
- remove stopwords and any tokens that are just empty strings

Example:

my_tokenizer("Until you have a dog you don't understand what could be eaten.")

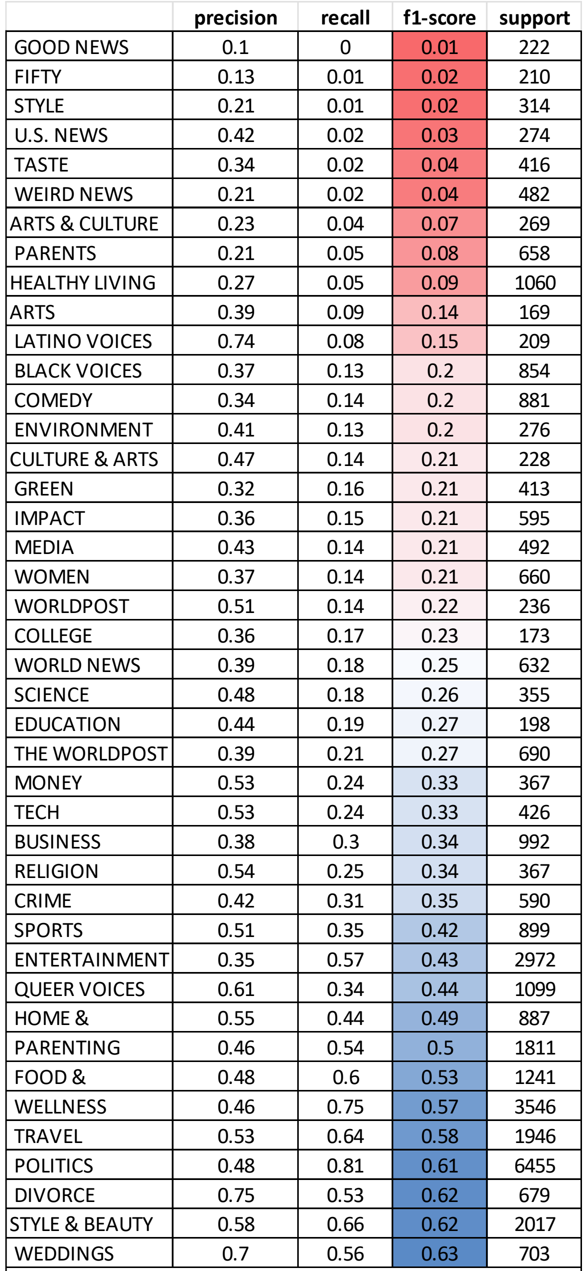
→ ['dog', 'dont', 'understand', 'could', 'eaten’]

Baseline Model

Logistic Regression:

Tokenizer	Pre-processing	Train	Test	Remarks
CountVectorizer	Custom tokenizer	61.40%	45.70%	Overfitting
	Custom tokenizer and upsampling "train set"	84.80%	36.80%	Severe Overfitting
TF-IDF	Custom tokenizer	52%	46.70%	**Promising!**
	Custom tokenizer and upsampling "train set"	66.50%	33.40%	Overfitting

Note: ‘Politics’ category constitutes about 1/6th of the total articles. Our model's test accuracy(46.7%) is better than random guessing which would be at max approx 16% (if one were to predict everything as Politics).



	precision	recall	f1-score	support
GOOD NEWS	0.1	0	0.01	222
FIFTY	0.13	0.01	0.02	210
STYLE	0.21	0.01	0.02	314
U.S. NEWS	0.42	0.02	0.03	274
TASTE	0.34	0.02	0.04	416
WEIRD NEWS	0.21	0.02	0.04	482
ARTS & CULTURE	0.23	0.04	0.07	269
PARENTS	0.21	0.05	0.08	658
HEALTHY LIVING	0.27	0.05	0.09	1060
ARTS	0.39	0.09	0.14	169
LATINO VOICES	0.74	0.08	0.15	209
BLACK VOICES	0.37	0.13	0.2	854
COMEDY	0.34	0.14	0.2	881
ENVIRONMENT	0.41	0.13	0.2	276
CULTURE & ARTS	0.47	0.14	0.21	228
GREEN	0.32	0.16	0.21	413
IMPACT	0.36	0.15	0.21	595
MEDIA	0.43	0.14	0.21	492
WOMEN	0.37	0.14	0.21	660
WORLDPOST	0.51	0.14	0.22	236
COLLEGE	0.36	0.17	0.23	173
WORLD NEWS	0.39	0.18	0.25	632
SCIENCE	0.48	0.18	0.26	355
EDUCATION	0.44	0.19	0.27	198
THE WORLDPOST	0.39	0.21	0.27	690
MONEY	0.53	0.24	0.33	367
TECH	0.53	0.24	0.33	426
BUSINESS	0.38	0.3	0.34	992
RELIGION	0.54	0.25	0.34	367
CRIME	0.42	0.31	0.35	590
SPORTS	0.51	0.35	0.42	899
ENTERTAINMENT	0.35	0.57	0.43	2972
QUEER VOICES	0.61	0.34	0.44	1099
HOME &	0.55	0.44	0.49	887
PARENTING	0.46	0.54	0.5	1811
FOOD &	0.48	0.6	0.53	1241
WELLNESS	0.46	0.75	0.57	3546
TRAVEL	0.53	0.64	0.58	1946
POLITICS	0.48	0.81	0.61	6455
DIVORCE	0.75	0.53	0.62	679
STYLE & BEAUTY	0.58	0.66	0.62	2017
WEDDINGS	0.7	0.56	0.63	703

Next steps…

- Refine pre-processing by finding ways to tokenize and remove noisy words/characters.

- Work with hyperparameters to see if the results get better.

- Optimize the parameters

- Explore other classification models such as SVM, Decision Trees, Naïve-Bayes

Thank You!

Appendix

File Name	Remarks	Train %	Test %	Status
ARK_V1	countvectorizer (custom tokenizer)		46	ok
ARK_V2	countvectorizer (custom tokenizer)		46	ok
	TFIDF vectorization		46	ok
ARK_V3	countvectorizer (custom tokenizer + remove html)	61	45	ok
	TFIDF vectorization	51	46	ok
ARK_V3_Sentence2Vec	Using the representation learning - (averaging the vector for the whole sentence).			WIP (re-run)
ARK_V4	upsampled			WIP (re-run)
ARK_V4_new_ngrams	upsampled + n-grams = 2	96	94	Mistake!
ARK_V5_upsampling	-	-	-	Error
ARK_V6_BERT	-	-	-	Error
ARK_V6_BERT_LM	BERT model (tensorflow)	42	44	ok
ARK_V7	custom tokenizer for TF-IDF and countvectorizer. Also, did upsampling for both tf-idf and countvec			

Learnings: Lots of back and forth in refining the features, model testing etc.