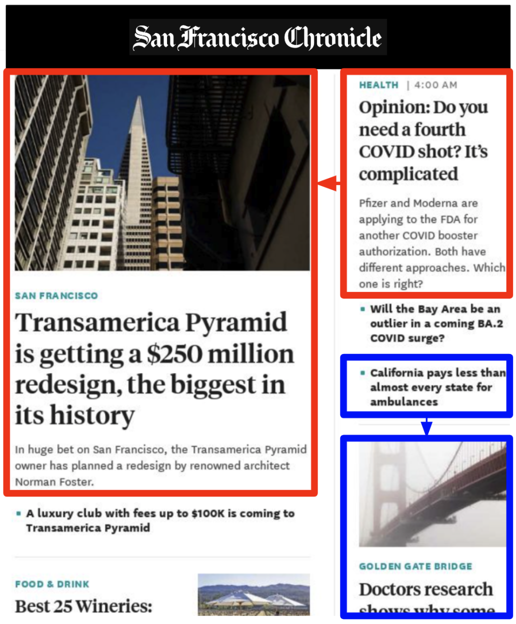
Spatial Layouts in News Homepages Capture Human Preferences

Alexander Spangher[1*], **Michael Vu**[2*], **Arda Kaz**[2*], **Naitian Zhou**[2], **and Ben Welsh**[3]

[1]Stanford University
[2]University of California, Berkeley
[3]Thomson Reuters
alexspan@stanford.edu

Abstract

Information prioritization plays an important role in the way we perceive and understand the world. Homepage layouts, which are daily and manually curated by expert human news editors, serve as a tangible proxy for this prioritization. In this work, we present NewsHomepages, a novel and massive dataset of over 3,000 news website homepages, including local, national, and topic-specific outlets, captured twice daily over a five-year period. We develop a scalable pairwise preference model to capture ranked preferences between news items and confirm that these preferences are stable and learnable: our models infer editorial preference with over 0.7 F1 score (based on human trials). To demonstrate the importance of these learned preferences, we (1) perform a novel analysis showing that outlets across the political spectrum share surprising preference agreements and (2) apply our models to rank-order a collection of local city council policies passed over a ten-year period in San Francisco, assessing their "newsworthiness". Our findings lay the groundwork for leveraging implicit cues to deepen our understanding of human informational preference.



San Francisco Chronicle

SAN FRANCISCO

Transamerica Pyramid is getting a $250 million redesign, the biggest in its history

In huge bet on San Francisco, the Transamerica Pyramid owner has planned a redesign by renowned architect Norman Foster.

- A luxury club with fees up to $100K is coming to Transamerica Pyramid

FOOD & DRINK

Best 25 Wineries:

HEALTH | 4:00 AM

Opinion: Do you need a fourth COVID shot? It's complicated

Pfizer and Moderna are applying to the FDA for another COVID booster authorization. Both have different approaches. Which one is right?

- Will the Bay Area be an outlier in a coming BA.2 COVID surge?

- California pays less than almost every state for ambulances

GOLDEN GATE BRIDGE

Doctors research shows why some

Figure 1: Signals that reveal editors' preference for one article over another. (1) **Position** (i.e. articles that are placed above, ↑, and left, ← are more important Hays (2018)). (2) **Size** (i.e. articles that are larger are more important) (3) **Graphics and Design** (i.e. articles with graphics and images). We release *NewsHomepages*, a large dataset of over 3,000 homepages, collected twice-daily over three years to study information prioritization in these signals *(we focus on (1) and (2))*. We show we can model these decisions and demonstrate the usefulness of these models on two downstream tasks.

1 Introduction

The way humans spatially organize information reflects a key signal of preference (Miller, 1956). The homepages of news organizations are one such artifact where spatial organization can be studied at scale: meticulously crafted by professional human editors, their layouts reflect the informational preferences of newspapers (Boukes et al., 2022).

A growing strain of research has leveraged weak signals from "found" data online (e.g. like Reddit upvotes) (Ouyang et al., 2022; Bai et al., 2022) to collect preference data for reinforcement learning from human feedback (RLHF) reward models. Given the importance of news in shaping our worldviews (McCombs, 1972), we find a lack of attention given to human preferences expressed through *news* spatial layouts.

To rectify this gap, we introduce *NewsHomepages*, a novel and massive dataset of homepage layouts, consisting of 363k homepage snapshots from over 3,000 news outlets spanning local, national, and topic-specific publishers, collected by a consortium of over 30 computer scientists, journalists and activists. We then ask two primary research questions: *How well can spatial layout signals be used to model editorial preferences? Do models for editorial preferences generalize across different*

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33404–33420
November 4-9, 2025 ©2025 Association for Computational Linguistics

corpora and are they be useful in different contexts?

To answer these questions, we first train a novel *layout parsing* model to better detect article positions on homepages, based on a novel bootstrapping approach (described further in Section 5). Next, we use this spatial information to infer weak judgments of preference: articles that *emphasized more*, based on their layout (examples shown in Figure 1), are preferred by editors over articles that are *emphasized less*. From these labels, we train pairwise preference models that *two articles predict which is preferred*, as a binary judgment (Section 6). By interpreting positional cues as indicators of preference, our preference models infer the relative importance of information and **achieve over 0.7 F1 score** in agreement with a human experiment.

Why is this interesting? What kinds of human preferences are revealed in these layouts? We demonstrate utility in two downstream experiments. In the first, we use preferences models trained on different homepages to *compare* the editorial preferences of these outlets: specifically, we compare how two different preference models agree on prioritization decisions. We find surprising nuances: for instance, despite *Fox News* (a right-leaning outlet) being topically dissimilar to *Mother Jones* (a left-leaning outlet), their newsworthy preferences are among the most correlated of outlets we studied. In the second, we used preference models trained on different outlets to rank-order local city council policies passed in San Francisco (Spangher et al., 2023). We show that these models capture a sense of "newsworthiness" valued by the different outlets: human evaluators judge policies ranked highly by these models as being particularly valuable for journalists and highly useful.

Contributions

- We introduce NewsHomepages, a massive dataset of 363k homepage layouts and develop parsing models to interpret layouts.
- We demonstrate that editorial preferences can be modeled. Our models achieve consistency across outlets and a high agreement with humans.
- We show via two case studies—(1) preference agreement between outlets, and (2) newsworthiness rankings for non-news corpora—that such models can generalize beyond the corpora we study and provide useful tools for analysis.

By furthering the study of human preferences in spatial layouts, we hope to open the door to many new avenues. We believe such work can facilitate greater cross-cultural preference comparisons in news, can help improve ranking algorithms online, can build more useful tools for journalists, and ultimately give us richer insights into the mechanisms influencing human perception.

2 Homepages Are a Source of Preference Signals

Visual cues for editorial preferences on homepages have a deep history in the design principles of physical newspapers (Barnhurst and Nerone, 2001). At *The New York Times*, for example, top editors and designers convened daily in the renowned *Page One* meeting (Usher, 2014) to determine the most important articles for the print newspaper the next day[1]. In the digital era, meetings like this evolved into *Homepage Meetings* (Sullivan, 2016), influencing the design and content placement on the website's homepage for the upcoming day. As such, homepages continue to be distillations of professional judgement and priorities.

One visual cue editors use is **positional placement**, with articles positioned towards the top and left of a page considered more important (Nielsen, 2006). This stems from observations that readers naturally begin scanning from the top-left corner (Bucher and Schumacher, 2006). Secondly, the **space** articles occupy is considered: larger articles or headlines are perceived as more important (García, 1987). In print media, prominence is conveyed through more column space; in digital media, longer headlines, featured images, and extended summaries. Finally, **graphics and design** also play a pivotal role in signaling the importance of news stories. Articles accompanied by photographs, videos, or other multimedia elements are often deemed more significant (Zillmann et al., 2001). The use of design elements (e.g. capital letters, bold fonts, and color) further enhances a story's prominence. In the rest of the paper, we introduce our dataset collection (§2) and processing pipeline (§5) before modeling placement decisions made by editors (§6) before downstream applications (Section 6).

[1] Terms like "above the fold" emerged to signal story-importance (i.e. the story is above the point at which the newspaper folds, so it is seen on newsstands)

3 Related Work

Prior research on *newsworthiness* underscores the role of editorial gatekeepers in selecting and emphasizing stories that align with certain values and organizational objectives (Galtung and Ruge, 1965; Harcup and O'Neill, 2001, 2017; Shoemaker, 1991; Herman and Chomsky, 2021). These studies highlight factors like timeliness and unexpectedness in shaping editorial choices, providing theoretical foundations for understanding news organizations' preferences. Our work extends these perspectives by examining how such judgments manifest in homepage layouts, offering a large-scale computational lens on editorial decision-making.

Visual cues, such as headline size, article position, and images, significantly influence how readers perceive the importance of news stories (Brooks and Pinson, 2022; Nass and Mason, 1990; Nielsen and Pernice, 2009; Bucher and Schumacher, 2006). Building on eye-tracking evidence that indicates a top-left viewing bias, researchers have proposed computational frameworks to model how layouts guide attention. We incorporate these findings by focusing on spatial arrangement as a measurable signal of prioritization, demonstrating that visual structure itself encodes editorial judgments.

Data-driven investigations into media bias and agenda-setting have traditionally centered on textual analysis (Gentzkow and Shapiro, 2010; Roberts et al., 2021; Misra, 2022; Silcock et al., 2024; Leetaru and Schrodt, 2013; Spangher et al., 2022), whereas emerging research leverages algorithms to support journalistic decision-making (Arya and Dwivedi, 2016; Diakopoulos et al., 2010). In contrast, our work examines how homepage layout attributes can be used to infer editorial preferences, linking presentation-driven signals with broader patterns of newsworthiness. This approach not only complements existing bias and recommendation studies but also opens pathways for new tools to help editors and developers refine news curation strategies.

4 Dataset Construction

4.1 Compilation of News Homepages

We compiled a list of 3,489 news homepages, as of the time of this writing, which we scraped twice daily, to capture morning and evening publishing cycles,[2] on an ongoing basis over a period of five years. From 2019-2024, we have collected a total of 363,340 total snapshots (details of each snapshot discussed in Section 4.2).

Our dataset collection is actively maintained and facilitated by a large contributing community of over 35 activists, developers and journalists. We collect homepages from national news outlets (e.g., *The New York Times*, *The Wall Street Journal*), state-level news outlets (e.g., *San Francisco Chronicle*, *Miami Herald*), as well as local and subject-matter-specific news sources. Table 1 provides a sample of the different categories of news homepages included in our dataset, and a full list can be found in the appendix. Additionally, we collect homepages from news websites of over 32 countries in 17 languages (please see Tables 6 and 8 for a more detailed breakdown). This is an ongoing and expanding effort: we encourage contributors to add their own news homepages of interest using for our suite of tools to scrape.[3] We hope to further diversify the news sources in the dataset that we collect.

4.2 Data Collection Pipeline

Our dataset collection runs in a `cron` job twice a day, and uploads data to Internet Archive. For each snapshot, we store the following information:

1. **All links on the page:** We store a flat-list of hyperlinks on every homepage and associated text.
2. **Full-page screenshots:** We store JPGs of each complete homepage as we render it.
3. **Complete HTML snapshots (subset of pages):** For a subset of homepages, we save a compressed version of the webpage, including all CSS files and images, using SingleFile[4].

In addition to our Internet Archive storage,[5] we also synchronize with Wayback Machine to store these homepages, providing a secondary backup and ensuring long-term preservation.

[2] We chose a twice-daily capture, every 12 hours, to capture morning and evening publishing cycles. This is historically when many news outlets will publish new articles and update homepages (Bergstrom, 2019).

[3] For more information on how to contribute, please see: `https://github.com/palewire/news-homepages`. For all code and data associated with this project, see `https://github.com/alex2awesome/homepage-newsworthiness-with-internet-archive`.

[4] `https://github.com/gildas-lormeau/SingleFile`, incidentally the same software that Zotero uses. In initial experimentation, we observed that capturing complete, compressed HTML snapshots was far more robust than capturing assets

[5] `https://archive.org/details/news-homepages`

Category	Example Outlets
National	The New York Times, The Wall Street Journal, NPR, Bloomberg
State-level	San Francisco Chronicle, Miami Herald, Chicago Tribune
Local	Sturgis-Journal, The Daily Jeffersonian, LAist, The Desert Sun
Subject-specific	The Weather Channel, Chessbase, ESPN
International	India Today, Ukrinform, BBC, Prensa Grafica, Japan Times

Table 1: *NewsHomepages* collects twice-daily snapshots from over 3,000 homepages across a wide breadth of *different* kinds of news outlets. Here we show several different categories of news, and show samples of different of news outlets in each category.

5 Dataset Processing

In order to robustly extract visual attributes for each article on a homepage (i.e. size, position, presence of graphics), we need to determine bounding boxes for all articles on a homepage. Examples of bounding boxes are shown in Figure 1: each bounding box, also referred to as *article card*, covers all information directly associated with that article. Layout parsing is a well-researched field (Shen et al., 2021; Li et al., 2020). However, homepages present unique challenges due to their diverse structures: text of varying size, fonts, colors and images are easily perceived by humans. Because none of the largest supervised datasets (Zhong et al., 2019) are specific to our task[6], we find that existing resources fail for parsing homepages. So, we bootstrap a supervised detection task.

5.1 Bootstrapping a Bounding Box Detector

Following other bootstrapping approaches (Amini et al., 2022), we: (1) develop a simple deterministic algorithm to generate candidate data, (2) apply a filtering step to exclude low-quality data, (3) use our high-precision dataset to train a more robust classifier. Figure 7, in the Appendix, provides an overview of the pipeline.

Step 1: Find Bounding Boxes Deterministically We design a deterministic algorithm, called the DOM-Tree algorithm, to start our bootstrapping process. At a high level, the algorithm traces each <a> tag in the Document Object Model (DOM) and extracts the largest subtree in the DOM that contains *only a single* <a> *tag* (illustrated in Figure 4, Appendix). This method can extract the maximal bounding box for each article, however it faces robustness challenges, for example, if a link exists *within* an article card (e.g. a link to an authors page, as shown in Figure 4b, Appendix.) We apply this algorithm to a subset of the NewsHomepages dataset, combining 15 homepages each from all outlets for which we have HTML files, JPEG snapshots, and hyperlink json files (approximately 15,000 homepages). Since each outlet typically maintains a consistent layout on their homepages across samples, we include more outlets for generalizability.

Step 2: Filter Low-Quality Bounding Box Extractions We take several filtering steps to prevent model degradation during bootstrapping (i.e. "drift" (Amini et al., 2022)). (1) First, we exclude non-news article links (e.g. log-in pages) by manually labeling 2,000 URLs as "news article" or "not" and training a simple text classifier[7]. The model achieves an accuracy of 96%. (2) Then, we exclude bounding boxes that did not overlap highly with link text. We determine this by first rendering the HTML pages as images and overlaying bounding boxes, then running OCR to extract the bounding-box text. (3) Finally, we exclude bounding boxes with improperly rendered images[8]. To address this, we again rendered HTML pages as an image and employed the YOLO object detection model (Redmon and Farhadi, 2018) to compare these images to the JPEGs in our archive. If a screenshot was not within 80% of the detection count of the archived snapshot, we discarded the snapshot. Overall, this multi-stage filtering process significantly reduced the number of boxes that did not correspond to actual articles and removed many websites that contained broken or corrupt data, enhancing the quality of the training data.

Step 3: Train a Robust Classifier Now, with our dataset in hand, we trained a Detectron2 model (Wu et al., 2019) to draw bounding boxes around article cards on pictures of homepages. Detection uses

[6] Existing work typically focus on parsing text around line-breaks (e.g. paragraph breaks). As can be seen in Figure 3, the same article box encompasses many line-breaks.

[7] The classifier a Logistic Regression classifier based off a bag-of-3-gram representation of each URL.

[8] Likely due to errors in HTML extraction or dead links

		FP#1	FP #2	FN #1	FN #2	Total Errors	% Correct
Challenge dataset	DOM-Tree algorithm	117	137	127	265	646	61.3%
	Detectron2 Model	25	23	27	87	162	90.3%
Clean dataset	DOM-Tree algorithm	12	20	0	13	45	97.1%
	Detectron2 Model	15	24	0	18	57	96.3%

Table 2: Error analysis of bounding box detection methods comparing the DOM-Tree algorithm and a Detectron2 model across two datasets: the challenge dataset and the clean dataset. The challenge dataset is formed by selecting the bottom 10% of articles based on the match between OCR-extracted text and retrieved link text (described in Section 5 Step 2), while the clean dataset contains well-matched articles. Error types are divided into false positives (FP #1: multiple articles in one box, FP #2: no articles in a box) and false negatives (FN #1: partially captured articles, FN #2: articles not captured). As can be seen, our trained model performs at par on the DOM-Tree algorithm in the clean settings and is far more robust in noisy settings.

ResNet-101 as a backbone with a Feature Pyramid Network (FPN) for extracting multi-scale features and Smooth L1 loss for bounding box regression. During training, we used a base learning rate of 0.02 with a linear warmup over the first 1000 steps. We trained the model for 10,000 steps with learning rate reductions after 5000 steps. A weight decay of 0.0001 and momentum of 0.9 were also employed. The training ran on 4×A40 GPUs for 24 hours.

5.2 Evaluation and Results

To evaluate the quality of our bounding box detection, we conducted manual validation for four types of errors: 1) bounding boxes that contain multiple articles, 2) bounding boxes that contain no articles, 3) bounding boxes missing parts of an article, and 4) articles that are not captured.

We used the OCR text-matching method, as described in Section 5, to identify particularly challenging homepages.[9] We compared errors on the cleanest 10% of homepages (Clean) and the least-clean 10% (Challenge). As shown in Table 2, our computer vision model (Detectron2) significantly improved the accuracy of bounding box detection in contexts where the DOM-Tree algorithm struggles (the Detectron2 model had a Card Correct % score of 90.3% while the DOM-Tree had a score of 61.3%). For Clean pages, the Detectron2 model performed similarly to the DOM-Tree algorithm, with error differences being minimal and both models achieving high accuracy (above 96%). The combination of deterministic algorithms and machine learning techniques allow us to achieve a more robust extraction of article attributes from diverse homepage layouts.

6 Preference Modeling

Given precise layout information for the 363k homepages in our dataset, we arrive at a core question of this research: can we model the editorial preferences expressed in homepage layouts?

6.1 Modeling Approach

Capturing preferences based on homepage placement presents a number of challenges. Firstly, publishing volumes are non-uniform: some days have lots of news (and many newsworthy stories) while others have less. Secondly, a homepage is intended to present a collection of articles as a cohesive bundle: individual articles do not exist in isolation (Tufte, 1990). Predicting the placement of a single article without considering surrounding context would limit information (Salganik et al., 2006); conversely, attempting to predict the placement of all articles simultaneously poses a combinatorial challenge. Finally, certain areas of homepages (e.g. "Latest News" feeds, which are ordered based on chronology) lack editorial decision-making altogether (Angèle, 2020). To address these challenges, we formulate our modeling task as a pairwise preference problem. Specifically, we consider pairs of articles (a_1, a_2) and train models to predict a binary preference variable p, where $p_o(a_1 > a_2) = 1$ if article a_1 is preferred over article a_2 for outlet o, and $p_o(a_1 > a_2) = 0$ otherwise.

Note: in the present work, we limit the layout variables we consider to: size and position. We explore three combinations of these variables to create weak labels for the preference variable, p:

Size-based Preference: We define $p_o(a_1 > a_2) = 1$ if article a_1 occupies more surface area on the homepage than article a_2, assuming prominent articles given more space (Lambert and Brock, 2005).

Position-based Preference: We set $p_o(a_1 > a_2) = 1$ if article a_1 is placed in a more favor-

[9] Figure 8, in the Appendix, demonstrates the resulting histogram for the distribution of OCR-match scores across our dataset.

	Size		Position x Size		Position	
Model Name	**F1** (Weak)	**F1** (Human)	**F1** (Weak)	**F1** (Human)	**F1** (Weak)	**F1** (Human)
Flan-t5-base	91.9	28.4	70.7	65.5	64.5	56.1
Flan-t5-Large	66.6	20.2	54.9	61.0	34.5	58.2
Roberta Base	91.0	26.6	64.9	62.9	37.3	53.9
Roberta Large	85.4	25.1	47.2	65.1	49.3	56.1
Distilbert-Base-Uncased	93.1	31.1	75.2	**70.4**	70.1	61.2

Table 3: F1 scores metrics on NYTimes data for different models. On the left, we show results in predicting the weak label, showing that coarser variables (e.g. size) tend to have greater consistency. On the right, we show comparison to a ground-truth preference ranking gathered via a human experiment: finer-grained variables (position x size) have the highest performance.

Outlet	Accuracy	F1	Recall	Prec.
phoenixluc	57.1	70.3	57.4	90.7
newsobserver	75.0	72.5	74.3	70.7
slate	72.4	61.6	66.2	57.7
jaxdotcom	75.2	63.4	65.5	61.4
arstechnica	64.7	17.5	41.4	11.1
airwaysmagazine	72.5	73.7	78.9	69.1
denverpost	73.7	67.8	70.5	65.4
thedailyclimate	82.0	80.9	81.3	80.6
breitbartnews	68.9	22.8	54.7	14.4
foxnews	67.3	38.6	55.6	29.5
motherjones	71.4	63.0	68.7	58.2
thehill	68.8	55.5	59.8	51.7
wsj	70.0	48.0	52.0	44.6

Table 4: Performance metrics on a sampling of outlets, including on ones we used for the downstream experiments in Section 7. Done with Distilbert-Base-Uncased model trained on position x size cues.

able location on the homepage than article a_2, such as higher up or more to the left, based on common reading patterns (Nielsen and Pernice, 2009).

Combined Size and Position Preference: Here, $p_o(a_1 > a_2) = 1$ if article a_1 either occupies more surface area or is in a more favorable position than article a_2, particularly focusing on articles in the top 10% by size on the page.

While there are other design variables that could give an even finer-grained preference (e.g. font, color, the presence of an image, as shown in Figure 1), we seek here to explore whether even a coarse weak labeling can still provide valuable insights. We leave exploration of *other* preference variables to future work. To model the weak preference labels discussed above, we train a simple Transformer-based binary classifier, `distilbert-base(X)` to classifies a text sequence X. We generate the text sequence X by concatenating input articles: `X=a`$_1$`<sep>a`$_2$ as input; the model learns to recognize the `<sep>` token as a boundary between the first and the second articles.

6.2 Modeling Variations

We explored modeling variations first on the *New York Times*[10]. We test 5 different models: {distilbert-base-uncased, flan-t5-base, flan-t5-large, roberta-base, roberta-large} and constructed a training dataset of 74,857 article-pairs and a test dataset consisting of 18,715 datapoints consisting of pairs of NYTimes articles from same homepages.

We observed exploding gradients in the flan-t5-large and RoBERTa-large models, motivating us to use a learning rate limit of 5e-5 for all the models and gradient clipping, for the sake of equal comparison. We applied Parameter-Efficient-Fine-Tuning (Mangrulkar et al., 2022) on flan-t5-base, flan-t5-large, roberta-base, roberta-large models to minimize overfitting, as we had limited of datapoints. We used 4xA40 GPUs and 16xA100 GPUs.

The distilbert-base-uncased model outperforms other models (Table 3) for our weak labels. We run a human validation experiment, enlisting a former New York Times journalist to rank-order 100 pairs of articles in our dataset. Weak labels agree with human preferences on >80% of pairs. Additionally, we recruited two more professional journalists (N=3 total) to annotate the same 100 NYT article pairs as pairwise preferences. Inter-annotator agreement was moderate (Fleiss' $\kappa = .6$; pairwise Cohen's $\kappa = \{.55 - .6\}$). The distilbert-base model trained on size×position achieved F1 = 0.70 against the aggregated human labels. See the Appendix for instructions, interface, and compensation details.

6.3 Dataset Selection and Processing

Next, our list of 3,000 outlets, we select 31 outlets for detailed analysis. We selected well-known

[10] We start with the *New York Times* as Spangher (2015) that meticulous rules, with full-time homepage editors hired, to that homepage layouts reflect preferences.

outlets in various categories, including different political leanings (left-leaning vs. right-leaning[11]), local and national levels, and varied subject matters such as science, chess and aviation. For each outlet, we collected between 200 and 300 homepage snapshots, resulting in 1,000 to 50,000 pairs of articles. We created an 80/20 train/test split and trained distilbert-base-uncased models for each outlet. We trained each model with 5e-5 learning rate limit, 3 epochs, 0.01 weight decay.

Each article in our dataset includes its textual representation as it appeared on the homepage. To enhance the reliability of our models, we undertake several data processing steps informed by preliminary experiments: (1) we only sample pairs of articles that are adjacent on the homepage, to curate preference pairs that are more likely to be challenging and topically similar. Secondly, we clean the textual representations by stripping out any times, dates, and formatting elements. We also remove author names to prevent the models from learning biases based on authors who might be favored by the organization. Please refer to Appendix A for a detailed list of the outlets used and the specific number of data points associated with each.

6.4 Results

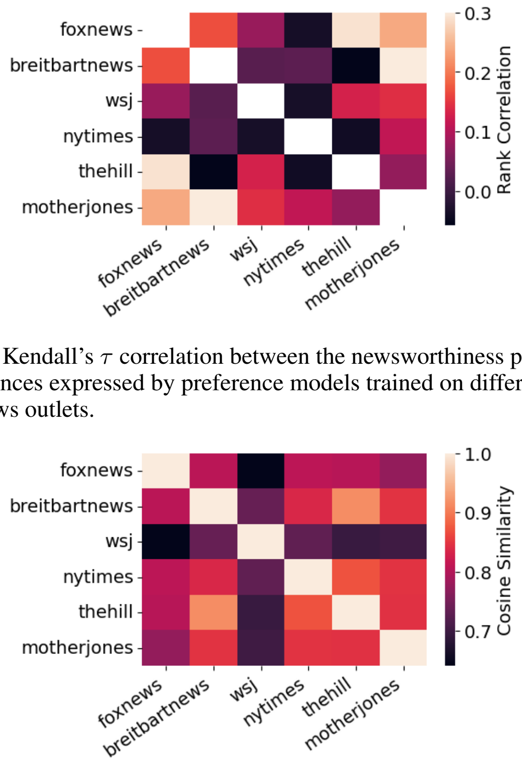
We show our results in Table 4. While some models (e.g. `Breitbart`) perform noticeably poorly, we note that the majority of our models score above $f_1 > .6$. We do not find a significant correlation between model performance and training set size. We were surprised to observe the tractability of this task; this indicates that many of the concerns we had about noise were either handled by our preprocessing steps, or not as important as we believed.

7 Demonstrations

To evaluate the practical utility of our models, we design two downstream tasks: (1) analyzing newsworthiness agreement between publishers, and (2) using newsworthiness models to rank corpora of interest to journalists.

7.1 Task 1: Newsworthiness Agreement Between Publishers

In this task, we aim to rank-order lists of news items drawn from a larger pool of articles to calculate the agreement rates for newsworthiness decisions between different news outlets. Previous research has observed surprising overlaps in sentiment and preferences between right-leaning and left-leaning outlets (Gentzkow and Shapiro, 2010), and we wish to quantitatively test this phenomenon using our preference models.



(a) Kendall's τ correlation between the newsworthiness preferences expressed by preference models trained on different news outlets.

(b) Cosine distance of average SBERT similarity between articles sampled from each outlet.

Figure 2: Comparison of Kendall's τ rank correlation (on newsworthiness judgements) and SBERT cosine similarity (on articles) across news outlets.

We selected 9 of the 31 outlets for which we trained preference models in the previous section. From each outlet, we sampled 1,000 articles, matching on variables such as topic, length, publication date, and other potential confounders. These 9 outlets were chosen because they represent a range of political viewpoints. For each model n_{o_i} (corresponding to outlet o_i), we used it to sort lists of 1,000 articles $\{a_1, a_2, \ldots, a_{1000}\}_{j=1}^{9}$ from outlets $\{o\}_{j=1}^{9}$. In other words, the output of applying model n_{o_i} to the article list from outlet o_j is a fully sorted list $n_{o_i}(A_j)$. We used the size × position model for this experiment, as performance was similar to the size-only model, and we believed that the multivariable models capture more newsworthiness information than the single-variable models.

We calculated Kendall's τ, a correlation measure for ordinal data, between each pair of sorted lists

[11] As classified by MediaBiasFactCheck.com

$(n_{o_i}(A_k), n_{o_j}(A_k))$ for all i, j, k, and averaged the correlations across j. Figure 2a shows the resulting correlation matrix. Some surprising insights emerge: *notably, Fox News, a right-leaning outlet, and Mother Jones, a left-leaning outlet, have one of the highest rates of agreement.*

Are these two outlets topically related? To ensure we are not merely measuring two outlets' *topical* similarity, we compared outlet-level embedding vectors. To derive these vectors, we sampled 100 articles per outlet and generated embeddings for each article using SBERT (Reimers and Gurevych, 2019). Then, we averaged these embeddings to create aggregated outlet-level embeddings (Sannigrahi et al., 2023). Our results, shown in Figure 2b, show that outlet-similarity, based off of article embeddings, aligns more closely with political differences: distinct right-wing clusters (e.g. *Fox News*, *Breitbart* and *Mother Jones*) segment from left-wing clusters (*New York Times*, *The Hill*, and *Mother Jones*). Taken together, these results suggest that newsworthiness preference is a novel and orthogonal variable to topical similarity.

7.2 Task 2: Surfacing Newsworthy Leads

In this task, we explore how well these newsworthiness judgments transfer outside of the news domain. In Spangher et al. (2023), the authors introduced the task of *newsworthiness prediction* as a detection task to aid journalists. They compiled a list of city council policies, which often serve as the basis for news stories, and trained binary classifiers to predict which policies would be *newsworthy enough* for journalists to write about. They demonstrated that these tools could find "needle-in-the-haystack" newsworthy policies (e.g. novel COVID policies on the eve of the pandemic) and substantively improve journalists' ability to find potential stories.

We suspect that editorial preference rankings learned from on news homepages can help us further identify newsworthy content, by identifying a more nuanced ranking of the *most* and *least* preferred stories of a news outlet. To test this hypothesis, we applied the preference models learned for each outlet to sort the list of the San Francisco Board of Supervisors' policies (compiled by (Spangher et al., 2023)). Then, we selected the top 10 items from the ordered lists n_{o_i} and used a large language model (LLM) to summarize the key points raised in each policy.[12]

The LLM's summarization results and examples are shown in Table 5. We observe various themes emerge, with subject-specific outlets like *The Weather Channel* highlighting policies related to environmental issues and *Fox News* highlighting policies related to public safety. We presented these results to a group of journalists ($n = 100$), and 81% of respondents indicated they were impressed and would consider using such a system in their workflow. These findings demonstrate the potential of our models to assist journalists in identifying newsworthy leads from large corpora of documents, thereby supporting timely reporting.

8 Discussion

Our novel dataset and experiments show that homepage editorial cues are normative, predictable decisions. We show that by modeling these decisions, we can provide a wealth of resources for (1) novel news analysis and (2) newsworthiness detection (Spangher et al., 2023; Diakopoulos et al., 2010). First, as we show in Section 7.1, editorial decision-making is distinct from simple topic preferences. In fact, information *prioritization* commonalities can be observed between outlets from vastly different political, social and topical backgrounds. Secondly, as we show in 7.2, preference models trained on news homepages can be transferred to related corpora (e.g. city council meeting minutes) and can surface relevant policies for journalists.

Both of these experiments reveal that editorial preferences be learned and transferred from the outlet's homepage to other domains, which led to novel analytical use-cases (e.g. comparing these preferences to *another* outlet's) and predictive use-cases (to use these preference models for rank-ordering leads). We see further applications for this direction of research and the use of these preference models, beyond what we explored, in: more extensive comparisons between outlet preferences (e.g. across global and cross-cultural divides) (Samir et al., 2024), assessing the *impacts* of different editorial preferences (e.g. on reader trust) (Ardèvol-Abreu and Gil de Zúñiga, 2017) and using learned preferences to improve algorithmic ranking systems on the web (Jia et al., 2024).

However, our results have to be taken with some important caveats. First, although we provided demonstrations to show that preferences could transfer across contexts, without *gold truth* about how an editor from *one outlet* would rank a set of

[12] We used GPT-4 for this experiment.

Outlet	Top Policies LLM Summaries	Examples of Policies
Weather Channel	Environmental Policies, Public Health and Emergency Response, Infrastructure and Development	Reducing nutrient pollution from wastewater; Accepting grants for forensic science improvements
Daily Climate	Environmental and Energy Policies, Urban Planning and Development	Agreement with North Star Solar; Building code enforcement
Fox News	Community and Public Safety Policy, Education and Social Policy, Fiscal and Economic Policy	Appointment of individuals to advisory committees; Appropriating funds for San Francisco Unified School District; Developing materials on domestic violence
Mother Jones	Social Policies, Environmental and Health Policies	Sanctuary City Protection; Urging Pardons; Edible Food Recovery and Organic Waste Collection
Ars Technica	Infrastructure Policies	System Impact Mitigation Agreement; 6th St. Substation
NYTimes	Social & Cultural Awareness Policies, Labor & Employment, Economic, Housing policies	Commemorative and Awareness Events; Labor Dispute Hearings; Affordable Housing Loans
WSJ	Economic and Infrastructure Policies, Governance and Legislative Policies	Contract modifications; Bond sales; Ground lease agreements; Charter amendments concerning commissions and departments related to aging and adult services

Table 5: Summaries of the top 10 most newsworthy policies published by the San Francisco Board of Supervisors, as ranked by models trained on 7 different homepages.

articles from *another outlet*, or *another domain*, we lack a *conclusive* measurement of transferability. While we experimented with many different ways of making our transfer more robust[13], our results in this direction were not conclusive. Secondly, while some preference models, when applied to city council policies, surfaced policies that were evaluated well by respondents, some seemed random. Further exploration is needed to determine why this was the case. Finally, despite the presence of non-English homepages in our dataset, we only tested with U.S.-based websites. It could be that spatial and layout preferences are different in different cultures. For example, in right-to-left writing systems, preferred positions might be different.

With these caveats in mind, we look forward to future work modeling spatial aspects of human preferences. We imagine a future where preferences learned from layouts can be applied more broadly to build tools for journalists, improve webpage layouts that are currently automated, and even understand more fundamental components of the human psyche.

9 Conclusion

This work introduces NewsHomepages, a large-scale dataset and modeling framework to study editorial prioritization through homepage layouts. The computational approach we develop for analyzing pairwise preferences gives us an operationalized new lens for learning from visual cues.

We leave to future work the incorporation of these cues into language modeling objectives. We hope this work inspires further exploration of how layout-based cues reflect the latent priorities of media organizations and how these cues can be leveraged for automated content curation, personalized news delivery, and broader social science inquiries into human attention and information design.

10 Acknowledgements

We would like to thank the large community of journalists, computer scientists and activists involved in curating these homepages: Mike Stucka, Alex Garcia, Davis Shaver, Jeremy B. Merrill and others. The first author would like to thank Bloomberg for funding a substantial part of this research through a four-year PhD fellowship (this research was conducted during his PhD at USC). Additionally, we would like to thank Jonathan May, Emilio Ferrara, Serdar Tumgoren and Cheryl Phillips for useful discussions throughout.

[13] For example, we attempted to train *additional* models to serve as in-domain and out-of-domain classifiers, and then ensembled them. In more detail, we trained a model to predict, for articles a_1 and a_2: $p_o(in\_domain|a_1, a_2) = 1$ whether a_1 belongs to outlet o while a_2 does not. Then, we multiplying the probabilities: $\hat{P}_o(a_1 > a_2) = p_o(in\_domain|a_1, a_2) \times p_o(a_1 > a_2)$

11 Limitations

This work, while advancing the study of editorial prioritization on homepages, comes with several limitations. First, the dataset, although large and diverse, predominantly focuses on English-language news outlets from the U.S., which may limit the generalizability of our models to international or non-English outlets. Despite the inclusion of some non-U.S. and non-English homepages, the models have not been explicitly evaluated on a broader range of languages or cultural contexts. This focus may overlook regional editorial conventions and biases that differ significantly from the U.S. context.

Another limitation is that the study focuses primarily on visual cues of newsworthiness, such as size, position, and graphical elements. While these cues are significant, they are not the only factors that influence editorial decisions. The models do not account for less visible but equally critical considerations, such as journalistic ethics, editorial mandates, or audience engagement metrics, which may influence homepage layouts but remain unquantified in this dataset.

Additionally, the weakly-supervised learning methods employed for layout parsing may struggle with more complex or irregular homepage designs. As described, the model had trouble generalizing to homepages with obscure HTML structures, leading to imperfect bounding box detections in some cases. This could result in misinterpretations of editorial significance, especially for websites with non-traditional or highly dynamic layouts.

Lastly, the results may have some bias due to the reliance on pairwise article comparisons. This method, while efficient, reduces the complexity of editorial decision-making into binary relationships, potentially overlooking more nuanced or multifaceted prioritization strategies that editors use in practice.

11.1 Societal Implications.

Preference models learned from homepage layouts can positively assist journalists by prioritizing promising leads within large document pools and by enabling cross-outlet comparisons that support media-studies analyses. At the same time, risks include potential homogenization of editorial style if outlets emulate “successful” rankings, gaming of layout signals, and misinference on atypical or non-standard layouts (e.g., dynamic pages or right-to-left designs). We recommend mitigations: transparent model cards and failure cases, outlet-specific calibration/caveats when transferring models across domains, and an opt-out mechanism for outlets that do not wish their layouts to be modeled.

11.2 Computational Budget

The computational resources required for this project were substantial. The model training involved 4×A40 GPUs for initial phases and 16×A100 GPUs for more extensive model fine-tuning and deployment. Training the custom Detectron2 model alone took 24 hours, with additional time required for fine-tuning and model testing across multiple outlets. While the experiments could be completed within this budget, more extensive experiments across all 3,000 outlets would require significantly more resources, especially when scaling to different languages and regional variations.

11.3 Use of Annotators

The dataset used in this work was primarily compiled automatically through web scraping, with minimal manual annotation. However, annotators were employed to manually label URLs to distinguish between news articles and non-news articles during the preprocessing phase. A set of 2,000 URLs was manually labeled, contributing to a more refined and accurate dataset. The authors performed this task themselves. However, beyond this, no human annotators were used to manually assess newsworthiness, as the models relied on position and size as proxies for editorial decisions.

References

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. 2022. Self-training: A survey. *arXiv preprint arXiv:2202.12040*.

Christin Angèle. 2020. Metrics at work: Journalism and the contested meaning of algorithms.

Alberto Ardèvol-Abreu and Homero Gil de Zúñiga. 2017. Effects of editorial media bias perception and media trust on the use of traditional, citizen, and social media news. *Journalism & mass communication quarterly*, 94(3):703–724.

Chandrakala Arya and Sanjay K Dwivedi. 2016. News web page classification using url content and structure attributes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 317–322. IEEE.

Yuntao Bai, Andy Jones, Keno Ndousse, Stanislav Fort, Amanda Askell, Sam Nisan, Anna Chen, Tom Conerly, Nelson Elhage, Gabriel Goh, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Kevin G. Barnhurst and John Nerone. 2001. *The Form of News: A History*. Guilford Press.

Guy Bergstrom. 2019. Understanding the newspaper news cycle. Accessed: 2025-05-19.

Mark Boukes, Natalie P Jones, and Rens Vliegenthart. 2022. Newsworthiness and story prominence: How the presence of news factors relates to upfront position and length of news stories. *Journalism*, 23(1):98–116.

Brian S Brooks and James L Pinson. 2022. *The Art of Editing: In the Age of Convergence*. Routledge.

Hans-Jürgen Bucher and Peter Schumacher. 2006. The relevance of attention for selecting news content. an eye-tracking study on attention patterns in the reception of print and online media.

Sasha Costanza-Chock and Pablo Rey-Mazon. 2016. Pageonex: New approaches to newspaper front page analysis. *International Journal of Communication*, 10:28.

Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122. IEEE.

Johan Galtung and Mari Holmboe Ruge. 1965. The structure of foreign news. *Journal of Peace Research*, 2(1):64–90.

Mario R. García. 1987. *Contemporary Newspaper Design: Shaping the News in the Digital Age*. Prentice Hall.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Tony Harcup and Deirdre O'Neill. 2001. What is news? galtung and ruge revisited. *Journalism Studies*, 2(2):261–280.

Tony Harcup and Deirdre O'Neill. 2017. What is news? news values revisited (again). *Journalism Studies*, 18(12):1470–1488.

Stephanie Hays. 2018. An analysis of design components of award-winning newspaper pages. *Elon Journal of Undergraduate Research in Communications*, 9(2):44–63.

Edward S Herman and Noam Chomsky. 2021. Manufacturing consent. In *Power and Inequality*, pages 198–206. Routledge.

Chenyan Jia, Michelle S Lam, Minh Chau Mai, Jeffrey T Hancock, and Michael S Bernstein. 2024. Embedding democratic values into social media ais via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–36.

Alan Lambert and Julie Brock. 2005. Layout complexity and visitors' attention on web pages: An eye-tracking study. *Journal of Digital Information*, 6(2).

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Minghao Li, Leyang Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Docbank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Maxwell E. McCombs. 1972. Agenda setting function of mass media. *Public Relations Review*, 3:89–95.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

Clifford Nass and Laurie Mason. 1990. On the study of technology and task: A variable-based approach. *Organizations and communication technology*, 46:67.

Jakob Nielsen. 2006. F-shaped pattern for reading web content. `https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/`. Accessed: 2023-10-06.

Jakob Nielsen and Kara Pernice. 2009. *Eyetracking Web Usability*. New Riders.

X. Ouyang, P. Wu, S. Ward, N. Joseph, G. Mishra, J. Hilton, K. Krawczuk, S. Wintle, T. Hanzlik, A. Askell, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Hal Roberts, Rahul Bhargava, Linas Valiukas, Dennis Jen, Momin M Malik, Cindy Sherman Bishop, Emily B Ndulue, Aashka Dave, Justin Clark, Bruce Etling, et al. 2021. Media cloud: Massive open source collection of global news on the open web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 1034–1045.

Matthew J. Salganik, Peter S. Dodds, and Duncan J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

Farhan Samir, Chan Young Park, Anjalie Field, Vered Shwartz, and Yulia Tsvetkov. 2024. Locating information gaps and narrative inconsistencies across languages: A case study of lgbt people portrayals on wikipedia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6747–6762.

Sonal Sannigrahi, Josef Van Genabith, and Cristina España-Bonet. 2023. Are the best multilingual document embeddings simply based on sentence embeddings? *arXiv preprint arXiv:2304.14796*.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Moser, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4528–4538. ACM.

Pamela J. Shoemaker. 1991. *Gatekeeping*. Sage Publications.

Emily Silcock, Abhishek Arora, Luca D'Amico-Wong, and Melissa Dell. 2024. Newswire: A large-scale structured database of a century of historical news. *arXiv preprint arXiv:2406.09490*.

Alexander Spangher. 2015. Building the next new york times recommendation engine. *The New York Times*, pages 08–26.

Alexander Spangher, Emilio Ferrara, Ben Welsh, Nanyun Peng, Serdar Tumgoren, and Jonathan May. 2023. Tracking the newsworthiness of public documents. *arXiv preprint arXiv:2311.09734*.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. NewsEdits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157, Seattle, United States. Association for Computational Linguistics.

Margaret Sullivan. 2016. The end of the page one meeting: Making way for the reader in choosing the news. *The New York Times*. https://publiceditor.blogs.nytimes.com/2016/03/16/the-end-of-the-page-one-meeting-making-way-for-the-reader-in-choosing-the-news/.

Edward R. Tufte. 1990. *Envisioning Information*. Graphics Press.

Nikki Usher. 2014. *Making news at the New York times*. University of Michigan Press.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xiaojie Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

Dolf Zillmann, Silvia Knobloch, and Zhao Yu. 2001. Effects of photographs on the selective reading of news reports. *Media Psychology*, 3(4):301–324.

Language	Count
English	975
Spanish, Castilian	44
Portuguese	36
Nepali	24
French	21
German	10
Japanese	9
Norwegian	8
Hindi	7
Hebrew	7
Russian	7
Italian	5
Ukrainian	5
Chinese	3
Afrikaans	3
Zulu	2
Xhosa	1

Table 6: Our corpus comprises homepages from 18 different languages. We assign each news outlet to the language of the majority of it's articles' languages (e.g. the *New York Times* sometimes publishes Spanish-language articles, but is predominantly and English-language newspaper).

A Dataset Details

In this section, we present more detailed dataset statistics. In Table 6, we show the different languages collected in our corpora and in Table 8 we show

'ainonline, airwaysmagazine, arstechnica, bleacherreport, breitbartnews, chessbase, cnet, denverpost, foxnews, jaxdotcom, jessicavalenti, jezebel, motherjones, newsobserver, nytimes, phoenixluc, rollcall, seattletimes, sfchronicle, sinow, slate, startelegram, studyfindsorg, thealligator, theathletic, thedailyclimate, thehill, weatherchannel, wired, wsj, yaledailynews'

	Domain		Position x Size	
Outlet	Train	Test	Train	Test
ainonline	2159	540	3844	962
airwaysmagazine	1233	309	1669	418
arstechnica	9349	2338	17883	4471
bleacherreport	3849	963	6689	1673
breitbartnews	7824	1957	15199	3800
chessbase	2094	524	3151	788
cnet	3769	943	6521	1631
denverpost	18802	4701	36607	9152
foxnews	62170	15543	125096	31274
jaxdotcom	4100	1026	7206	1802
jessicavalenti	455	114	512	129
jezebel	6270	1568	10956	2740
motherjones	2443	611	3572	893
newsobserver	7538	1885	14078	3520
nytimes	38432	9608	74857	18715
phoenixluc	209	53	305	77
rollcall	6572	1643	12436	3109
seattletimes	20942	5236	40882	10221
sfchronicle	5600	1401	10952	2739
sinow	176	44	315	79
slate	23527	5882	45470	11368
startelegram	11964	2992	22932	5734
studyfindsorg	450	113	403	101
thealligator	3046	762	5432	1359
theathletic	22722	5681	44463	11116
thedailyclimate	7211	1803	13688	3423
thehill	30800	7700	59965	14992
weatherchannel	4551	1138	8094	2024
wired	2058	515	3196	800
wsj	15569	3893	30496	7625
yaledailynews	1378	345	2527	632

Table 7: Size of training and test sets, in terms of # of pairs, used in our experiments.

B News Homepage Layouts

In Figure 3, we show several areas of a homepage where editorial policies are likely to be unclear and challenging to model. In the top section, a "Breaking News" feed shows articles shortly after they are published. They usually do not stay long in these positions (Costanza-Chock and Rey-Mazon, 2016), so there is high variability in this section. In the middle section, a "Section Fronts" show top articles in each section, each combining the different priorities of the desks (Angèle, 2020). Finally, in this reporter's experience, the bottom of a homepage was affectionately called the "Gutter". However, it is more commonly referred to as a "footer"[14]

We can extend this analysis by visualizing the flow of articles on a homepage over time. We show in Figure 5 where articles tend to get added and deleted overall, as well as where they stay the longest. In Figure 6, we show how articles shift frequently around a homepage.

[14] https://alyamanalhayekdesign.com/blog/the-parts-of-a-webpage-a-complete-list/

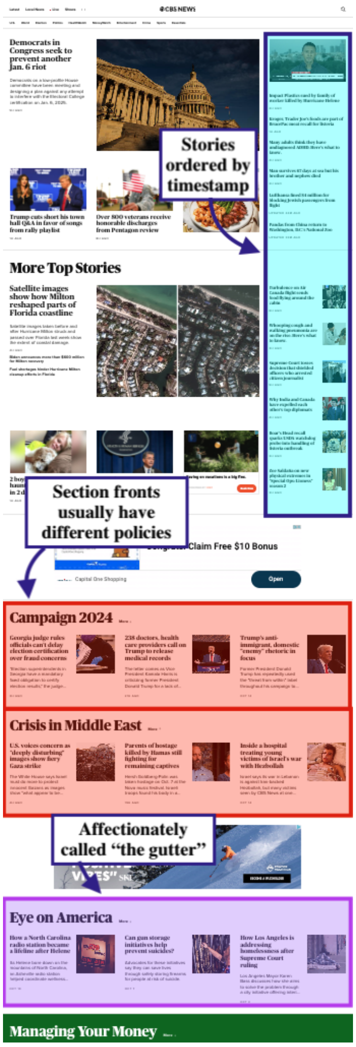
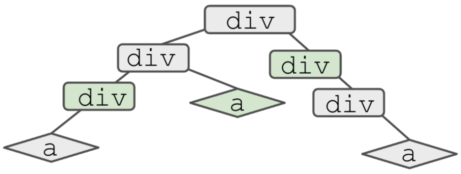

Stories ordered by timestamp
More Top Stories
Section fronts usually have different policies
Campaign 2024
Crisis in Middle East
Affectionately called "the gutter"
Eye on America
Managing Your Money

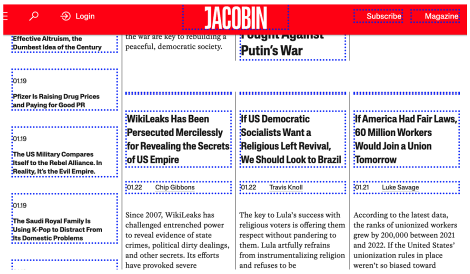
Figure 3: We show three sections of a sample homepage (from CBS News) where editorial decisions for different reasons. We highlight the "Breaking News" Section, "Section Fronts" and "The Footer".

Country	Count
United States	892
Brazil	37
United Kingdom	32
Nepal	24
Canada	20
South Africa	18
France	17
Spain	13
Mexico	13
India	10
Japan	9
Argentina	9
Israel	9
Germany	9
Russia	8
Norway	8
Ukraine	6
Ireland	6
Italy	5
New Zealand	4
Austria	3
Taiwan	3
Colombia	2
Australia	2
Uruguay	1
Qatar	1
Belgium	1
Latvia	1
Bosnia and Herzegovina	1
Georgia	1
El Salvador	1
Lebanon	1

Table 8: Countries of origin for the homepages we collect, based on where the organization is based.



(a) Our deterministic algorithm starts at all $\langle a \rangle$ nodes and recursively traverses up the DOM to find maximal subtrees with one $\langle a \rangle$. Green nodes shown are article bounding boxes.



Effective Altruism, the Dumbest Idea of the Century

01.19

Pfizer Is Raising Drug Prices and Paying for Good PR

01.19

The US Military Compares Itself to the Rebel Alliance. In Reality, It's the Evil Empire.

01.19

The Saudi Royal Family Is Using K-Pop to Distract From Its Domestic Problems

the war are key to rebuilding a peaceful, democratic society.

WikiLeaks Has Been Persecuted Mercilessly for Revealing the Secrets of US Empire

01.22 Chip Gibbons

Since 2007, WikiLeaks has challenged entrenched power to reveal evidence of state crimes, political dirty dealings, and other secrets. Its efforts have provoked severe

Fought Against Putin's War

If US Democratic Socialists Want a Religious Left Revival, We Should Look to Brazil

01.22 Travis Knoll

The key to Lula's success with religious voters is offering them respect without pandering to them. Lula artfully refrains from instrumentalizing religion and refuses to be

If America Had Fair Laws, 60 Million Workers Would Join a Union Tomorrow

01.21 Luke Savage

According to the latest data, the ranks of unionized workers grew by 200,000 between 2021 and 2022. If the United States' unionization rules in place weren't so biased toward

(b) Failure cases (missing text area) with the deterministic algorithm.

Figure 4: Illustration of our deterministic bootstrapping algorithm and a failure case. Here, when non-article links exist, we misunderstand the full area of an article, excluding the text below.

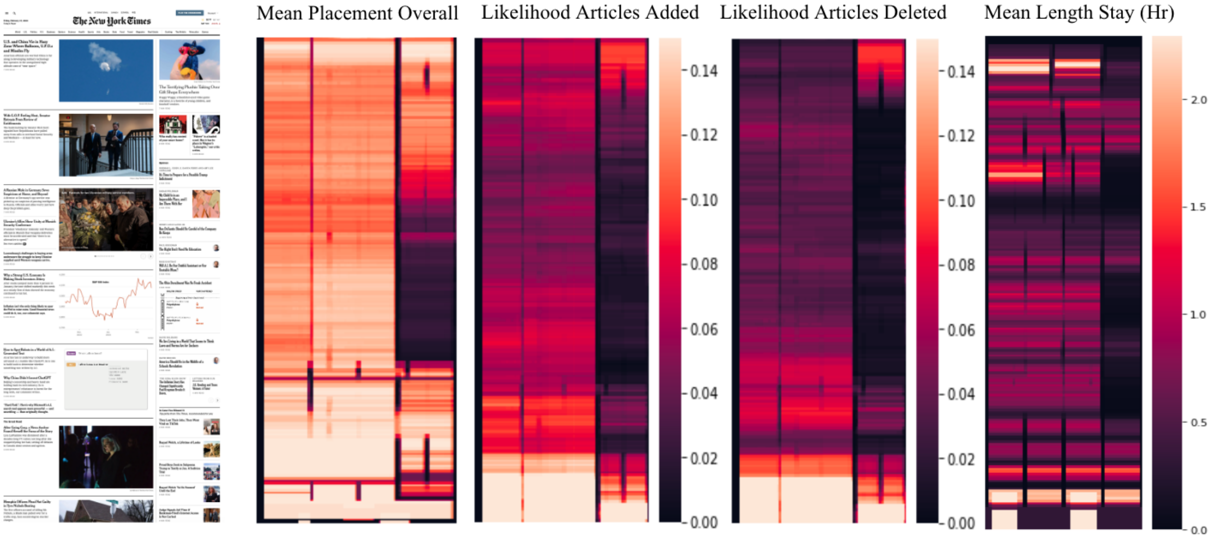


Figure 5: Different analyses we run on bounding boxes across time: average locations of bounding boxes on a homepage, locations where articles are added first, locations where they are removed, and the average time articles in various locations spend.

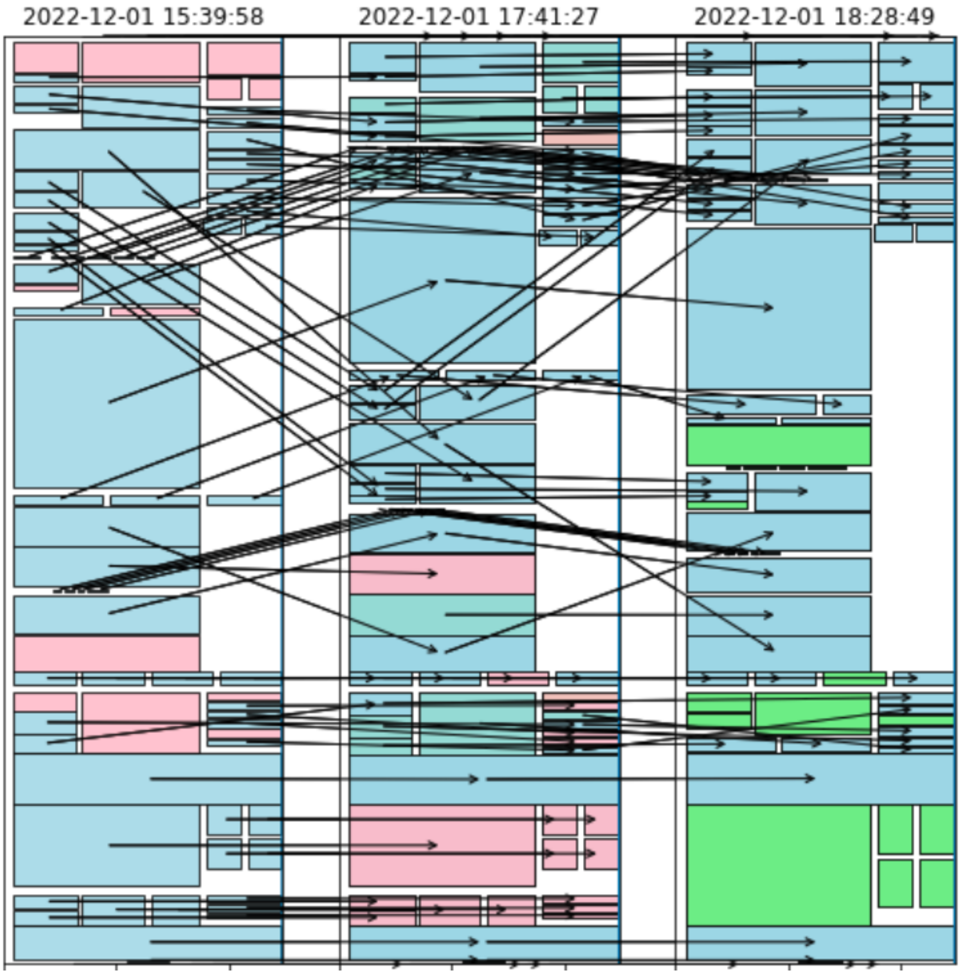


Figure 6: With our suite of tools for parsing homepages, we can examine on a granular level the movement of an article across the homepage.

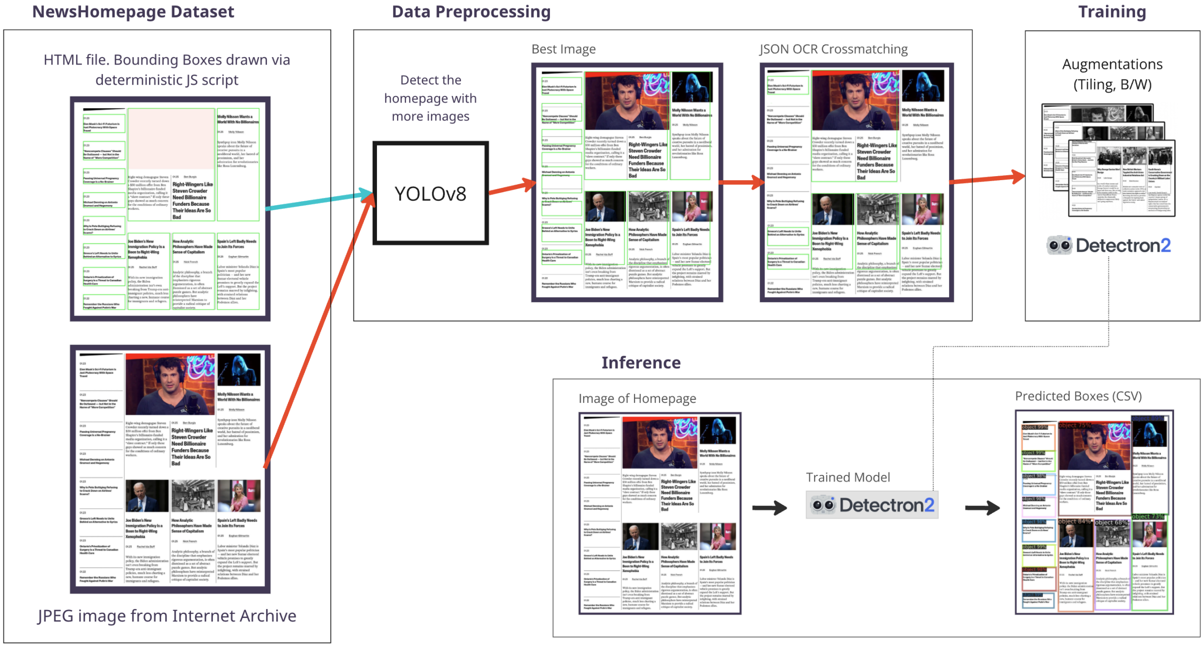


Figure 7: This diagram is an overview of the data preparation and training of the Detectron2 model for predicting bounding box on websites.

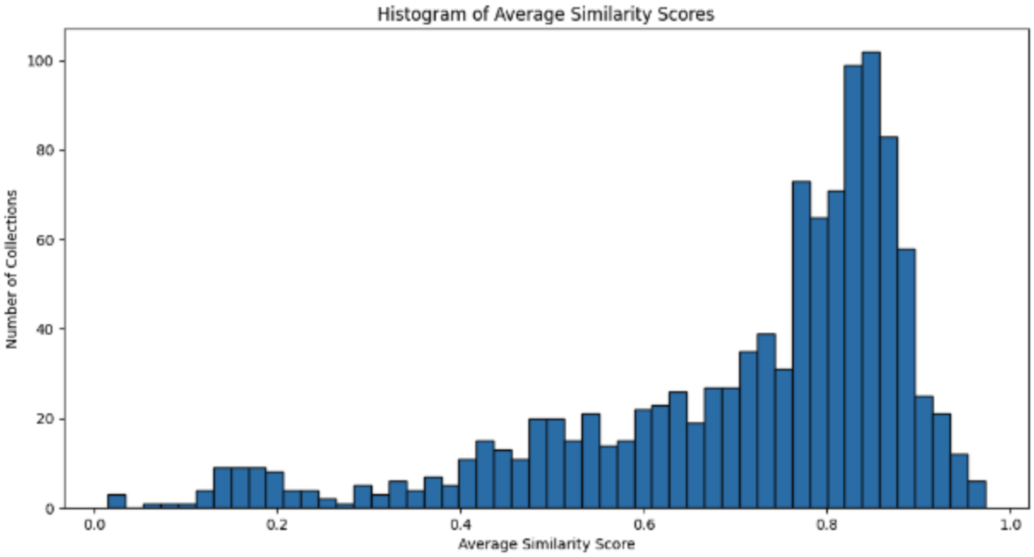


Figure 8: When sorting our sources to determine the ones most difficult for the DOM-Tree algorithm, we define the Average Similarity score to be a general measure as to how well the bounding box's text match the article's JSON file containing text/link pairs. High similarity score means high bounding box accuracy, and vice versa.