NewsHomepages: Homepage Layouts Capture Information Prioritization Decisions

Anonymous ACL submission

Abstract

The prioritization of information plays a pivotal role in how humans perceive and understand the world. Homepage layouts serve as a tangible proxy for this prioritization. In this work, we present NewsHomepages, a large dataset of over 3,000 websites (including local, national and topic-specific homepages) captured twice daily over a three-year period. We develop models to perform pairwise comparisons between news items to infer their relative significance, showing over 0.8 F1 score across the majority of tested cases. To illustrate that modeling organizational hierarchies has broader implications, we applied 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 organizational cues to deepen our understanding of information prioritization.

1 Introduction

The way humans prioritize and organize information is central to attention and understanding. Prior work has explored the need for effective prioritization to manage information overload (Miller, 1956) and including case-studies of prioritization approaches (Rosenfeld, 2002; Hays, 2018). However, larger scale analyses, specifically with an aim towards predictive modeling, are often limited by lack of available data and methods.

In this work, we seek to rectify this by focusing on a prominent source of publicly available, highly curated examples of prioritization decisions. News organizations' homepages are meticulously crafted by professional human editors: their layouts offer manifestations of the news organization's prioritization principles (Boukes et al., 2022). We use homepage layouts to study information organization at scale. We present *NewsHomepages*, to our knowledge the first large-scale homepage layout dataset, consisting of over 363k full page screenshots, HTML, extract links and other metadata from over 3,000 websites, which we captured twice daily over a three-year period, from local, magazine, and national news homepages. With this dataset, we ask the question: *how well can we model the editorial decisions behind homepage layout? What do these learned decisions tell us about the priorities of the newspaper and the world around us?*

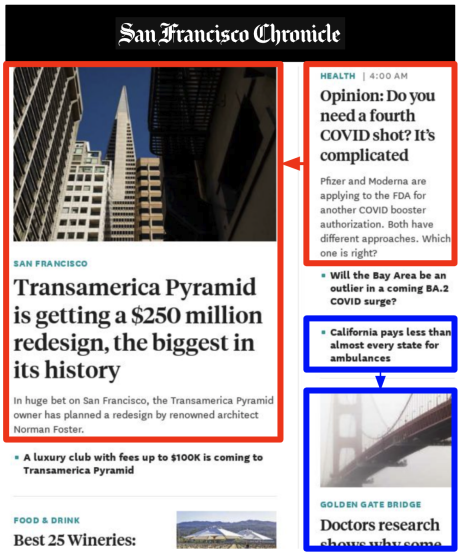


Figure 1: Shown above are two signals of importance that professional editors apply to guide reader attention, according to Hays (2018). (1) **Position** (i.e. articles that are placed above, ↑, and left, ←). (2) **Size** (3) **Graphics and Font**. We release *NewsHomepages*, a large dataset of over 3,000 homepages, collected twice-daily over three years, study information prioritization in this setting. We show can model these decisions at scale and demonstrate the usefulness of these models on an established downstream task.

To answer these questions, we develop models to process and learn from our dataset at scale. First, we develop a novel weakly-supervised approach to robustly parse images of homepages homepages

and determine precise spatial alignments of news articles. Then, we train models to perform pairwise comparisons between articles based on their positions within layouts. By interpreting positional cues as indicators of significance, our models infer the relative importance of information and achieve over 0.8 F1 score in most scenarios we tested.

The ability to learn these decisions with high accuracy indicates the robust signal this dataset captures. We conducted two downstream experiments utilizing these models; in each, we applied our pairwise comparison models to rank-order a list of input texts. In the first experiment, we use models trained on the newsworthy judgements of one news outlet to sort the news articles of another. We find surprising nuances: for instance, despite *Brietbart* (a right-leaning outlet) being topically dissilimar to *Mother Jones* (a left-leaning outlet), their newsworthy preferences are among the most correlated of outlets we studied. In the second analysis, we used each outlets newsworthy judgements to rank-order local city council policies passed in San Francisco (Spangher et al., 2023). By assessing these policies' "newsworthiness" through the lens of organizational prominence, we highlight how implicit cues in information structure can transfer between domains and develop concrete tools (in this case, to help human journalists find relevant story-topics).

Our contributions are as follows:

- We introduce NewsHomepages, large-scale dataset of hompage layouts from over 3,000 local, national and topic-specific news organizations, and develop a robust, weakly-supervised method to parse articles.

- We demonstrate that editorial principles behind information prioritization can be learned, by training pairwise comparison models to predict article size and positionality. Across all domains that we test, our models achieve $> .8$ f1-score.

- We show via two case-studies – (1) newsworthiness agreement between outlets, and (2) generating newsworthiness rankings for non-news corpora – that such learned principles can generalize beyond the corpora we study and provide useful tools for end-users.

This work opens new avenues for exploring how implicit cues in digital environments reflect latent priorities and organizational principles. By modeling these cues, we can gain richer insights into the mechanisms that influence human perception.

2 Background

Visual cues for editorial importance on homepages have a deep history in the design principles of physical newspapers (Barnhurst and Nerone, 2001) and result from deliberate editorial processes. 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 take 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 capital letters, bold fonts, and color further enhances a story's prominence.

Despite well-established practices cemented over decades, we find few large-scale computational analyses studying these attributes. Our work aims to enable a more precise study of editorial judgement tracking editorial cues in homepage layouts over a large corpus. In the next section, we will introduce our data collection processes. In the rest of the paper, we will introduce our dataset collection and processing (Section 3 and 4). Then, we will introduce the modeling approaches we took to capture placement decisions made by editors and their downstream applications (Section 5).

[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 Dataset Collection

3.1 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 on an ongoing basis over a period of 3 years. From 2019-2024 we have collected a total of 363,340 total snapshots. 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.[2] We hope to further diversify the news sources in the dataset that we collect.

3.2 Data Collection Pipeline

Our dataset collection runs in a `chron` 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 the text associated with them.

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[3]. In initial experimentation, we observed that capturing complete, compressed HTML snapshots was far more robust than capturing assets.

In addition to our Internet Archive storage,[4] we also synchronize with Wayback Machine to store these homepages, providing a secondary backup and ensuring long-term preservation.

4 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. The extraction process must be generalizable across various colors, fonts, occlusions, and other variations commonly found in web layouts.

4.1 Challenges in Layout Parsing

Layout parsing is a well-researched field, with applications ranging from document analysis to web page understanding, and tools for a variety of domains (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 understood by humans to describe the same article, yet, because none of the largest supervised datasets (Zhong et al., 2019) are specific to our task, we find that existing resources cannot meet our use-case. Thus, we bootstrap a supervised detection, which we describe next.

4.2 Bootstrapping a Bounding Box Detector

The approach we take to simultaneously collect a training dataset and train a model is a bootstrapping approach (Amini et al., 2022): first, we apply a simple heuristic to generate candidate data, then we apply a filtering step to exclude low-quality data. Finally, we use our high-precision dataset to train a classifier. Our approach combines deterministic algorithms with machine learning techniques to improve robustness and generalization across different homepage layouts. Figure 7 provides an overview of the entire data preparation and training pipeline.

Step 1: Constructing a Train/Validation Dataset In order to construct our bootstrapping dataset, we sample a subset of the NewsHomepages dataset combining 15 homepages from all outlets (containing HTML files, JPEG snapshots, and hyperlink json files). We do this in order to construct a balanced dataset across a wider array of styles. We also assume that every source typically maintains a consistent layout, so more data from a wider variety of sources is preferrably for generalizability. This gave us approximately 15,000 high-resolution images (ranging from 1k x 10k to 1k x 10k-60k

[2] For more information on how to contribute, please see: [WITH-HELD]

[3] `https://github.com/gildas-lormeau/SingleFile`, incidentally the same software that Zotero uses

[4] `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: Sample of News Homepages by Category

pixels), which were full-page screenshots of news homepage websites.

Step 2: Deterministic Bounding Box Extraction We design a deterministic algorithm, called the DOM-Tree algorithm, to start our bootstrapping process. The goal of our algorithm is to extract the maximal bounding box for an article card, based on the HTML structure of the page. As illustrated in Figure 4, the algorithm seeks to extract the largest area of the Document Object Model (DOM) that encompasses a single link in its hierarchy. This method provides a large bounding box around an article but faces robustness challenges, as shown in Figure 4b. As can be seen in the image, links such as author bio pages or other non-article induce the algorithm to miss significant portions of the article card. Despite these difficulties, we find that the DOM-Tree algorithm manages to perform near perfect on the majority of homepages, but struggle significantly on a minority of websites with more obscure HTML structures. Our goal with building a computer-vision based model is to generalize the performance of this deterministic method to all website layouts.

This motivates our search for a more robust learning method. We take several filtering steps to purify our extracted bad data and prevent "drift". First, we excluded non-news article links (e.g. log-in pages) by training simple text classifier to distinguish between URLs to news articles and others. We manually labeling over 2,000 URLs, achieving an accuracy of 96%. Then, we excluded bounding boxes that did not overlap highly with link text. We determined this by first rendering the HTML pages as images and overlaying bounding boxes, then running OCR to extract the bounding-box text. Finally, a third filtering approach we took was to exclude bounding boxes with improperly rendered images, likely due to errors in HTML extraction or dead links. 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, thereby enhancing the quality of the training data.

Step 3: Detectron2 Model Training Now with our dataset in hand, we trained a custom Detectron2 model, based on (Wu et al., 2019). This model tiles images to manage their size (bounded between 1k x 10k–60k pixels) and the configuration included several key parameters: (1) Backbone: ResNet-101 with a Feature Pyramid Network (FPN) for extracting multi-scale features.(2) Bounding box regression: The model used Smooth L1 loss for bounding box regression with predefined weights to account for the size variability of the article cards. (3) Learning rate and optimization: A base learning rate of 0.02 was used with a linear warmup over the first 1000 iterations. The model underwent 10,000 iterations in total, with learning rate reductions after 5000 iterations. A weight decay of 0.0001 and momentum of 0.9 were also employed. The training ran on 4x A40 GPUs for 24 hours. The final model output, including weights and evaluation metrics, was saved and used for further analysis.

4.3 Evaluation and Results

To evaluate the quality of our bounding box detection, we conducted manual validation on the basis of two questions:

(1) How does our model perform compared to the DOM-Tree algorithm on a set of articles where the DOM-Tree algorithm demonstrates **high** accuracy?

(2) How does our model perform when applied to a set of articles where the DOM-Tree algorithm performs **poorly**?

		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 4 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, the our trained model performs at par on the DOM-Tree algorithm in the clean settings and is far more robust in noisy settings.

We manually validated four types of errors: false positives (FP) (#1: Bounding boxes containing more than one article and #2: Bounding boxes that do not contain any articles) and false negatives (#1 Bounding boxes that miss parts of the article and #2: Articles that are entirely not captured).

We began by assessing the performance of the DOM-Tree algorithm, using the OCR text-matching method described above on all homepages in our dataset. Figure 8 demonstrates the resulting histogram for the distribution of OCR-match scores across our dataset. We created two datasets for comparison: the top 10% of homepages (Clean) and the bottom 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 relative error difference between the two models was notable in this context, as the CV-based Detectron2 model had an Card Correct % score of 90.3% from the DOM-Tree algorithm's 61.3%. For the top-performing sources, 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 allowed us to achieve a more robust extraction of article attributes from diverse homepage layouts. By improving the accuracy of bounding box detection, our model enables precise measurement of article prominence based on size, position, and graphical elements.

5 Homepage Modeling

Given precise layout information for the 363k homepages in our dataset, we arrive again at the core question of this research: how regular and predictable is an outlet's layout decisions? We initially hypothesized that modeling homepage placement would be challenging. As shown in Figures 3, 5, 6[5] and described by (Angèle, 2020), certain areas of many homepages lack clear editorial consistency. This introduces noise and makes it difficult to learn a uniform policy guiding article placement decisions. Further, learning a single set of policies is challenged by the changing news cycle; some days have lots of news while others have less.

5.1 Modeling Approach

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 the context of other articles would be overly noisy and potentially ineffective (Salganik et al., 2006). Conversely, attempting to predict the placement of all articles simultaneously poses a combinatorial challenge that is computationally infeasible.

To address this issue, 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 $p(a_1 > a_2) = 1$ if article a_1 is preferred over article a_2, and $p(a_1 > a_2) = 0$ otherwise.

We explore three variations of preference criteria for the preference variable, p:

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

2. **Position-based Preference**: We set $p(a_1 > a_2) = 1$ if article a_1 is placed in a more favorable location on the homepage than article a_2, such as higher up or more to the left, based

[5] In Appendix B

	Size			Position x Size		
Model Name	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Flan-t5-base	91.9	91.8	92.0	70.7	80.5	63.1
Flan-t5-Large	66.6	51.1	95.4	54.9	43.4	74.7
Roberta Base	91.0	91.7	90.4	64.9	78.1	55.6
Roberta Large	85.4	85.9	84.9	47.2	74.4	34.6
Distilbert-Base-Uncased	93.1	93.3	92.9	75.2	81.8	69.6

Table 3: Performance metrics on NYTimes data for different models

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 6.

on common reading patterns (Nielsen and Pernice, 2009).

3. **Combined Size and Position Preference**: Here, $p(a_1 > a_2) = 1$ if article a_1 either occupies more surface area or is in a more favorable position than article a_2, particularly focusing on articles that are in the top 10

To model these preference variables, p, we train a simple Transformer-based binary classifier, `distilbert-base(X)`, which classifies a text sequence X. Our model concatenates the input articles: `X=a`$_1$`<sep>a`$_2$ as input; the model learns to recognize the `<s>` token as a boundary between the first and the second articles.

5.2 Modeling Variations

We explore different modeling variations on the *New York Times* homepages, as they have a variety of content, shares and functionalities on their site (Spangher, 2015). 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 experienced exploding gradient in flan-t5-large, Roberta-large model, motivating us to use a learning rate limit of 5e-5 for all the models, 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 amount of datapoints.

As can be seen in Table 3, distilbert-base-uncased model outperforms other models, so we choose this model. We train on 4xA40 GPUs and 16xA100 GPUs and implement gradient clipping after observing gradient explosion.

5.3 Dataset Selection and Processing

From our list of 3,000 outlets, we select 31 outlets for detailed analysis. We selected well-known outlets in various categories, including different political leanings (left-leaning vs. right-leaning[6]), 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]As classified by MediaBiasFactCheck.com

5.4 Results

We show our results in Table 4. While some models (e.g. `Brietbart`) perform noticably 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 such high results; 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.

6 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.

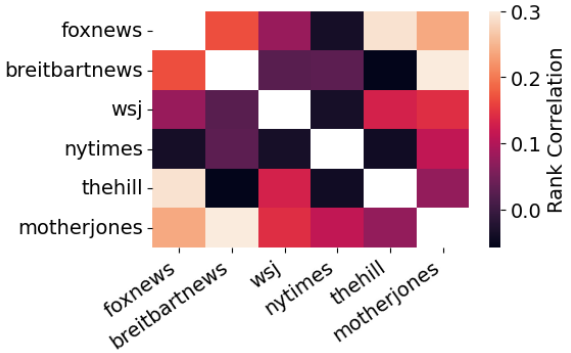
6.1 Task 1: Newsworthiness Agreement Between Publishers

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.

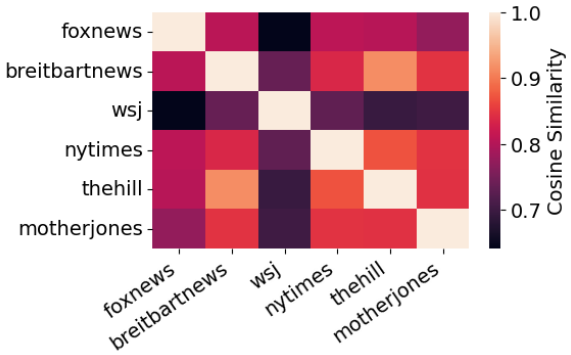
We selected 9 of the 31 outlets for which we trained preference models in the previous section. From each outlet, we sampled 1,000 articles, matching on variables such as topic, length, publication date, and other potential confounders. These 9 outlets were chosen because they represent a range of political 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 chose the size $\times$ position model for this experiment, as performance was similar to the size-only model, and we believed that the multi-variable models capture more newsworthiness information than the single-variable models.

We then calculated Kendall's τ, a correlation measure for ordinal data, between each pair of sorted lists $(n_{o_i}(A_k), n_{o_j}(A_k))$ for all i, j, k, and averaged the correlations across j. The resulting correlation matrix is displayed in Figure 2a. Some surprising insights emerge from this analysis. Notably, *Breitbart*, a right-leaning outlet, and *Mother Jones*, a left-leaning outlet, have one of the highest rates of agreement.



(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.

To establish a baseline and ensure we are not merely capturing topic overlap (despite matching on topics), we conducted a simple SBERT embedding experiment (Reimers and Gurevych, 2019). We sampled a set of 100 articles per outlet, generated embeddings using SBERT, and averaged these embeddings to create single outlet-level embeddings, as shown in Figure 2b. These embedding-level similarities align more closely with topical overlaps, indicating distinct right-wing and left-wing clusters with some overlap in between.

Taken together, these results suggest that newsworthiness preference offers a compelling and orthogonal variable for study, beyond mere topical similarity.

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.

6.2 Task 2: Surfacing Potentially Newsworthy Leads

In this task, we explore how well these newsworthiness judgments transfer outside of the news domain. Building on the work of Spangher et al. (2023) on newsworthiness prediction for lead generation, we utilize their list of San Francisco Board of Supervisors' policies. They studied newsworthiness prediction as a detection and alerting system for journalists.

We applied the models from each outlet to sort the list of Board of Supervisors' policies. 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.[7]

The LLM's summarization results and examples are shown in Table 5. An emergent theme appears, with subject-specific outlets like *The Weather Channel* highlighting policies related to environmental issues. We presented these results to a group of journalists, 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 investigative journalism and timely reporting.

7 Discussion

Our demonstrations show two core findings: (1) editorial priorities and decision-making can be inferred simply by examining the layout decisions made on homepages. This decision-making is distinct from simple topic preferences, as we show. In fact, commonalities about decision-making can be observed between outlets that appear distinct topically. (2) Newsworthiness judgements can transfer to non-news domains. To make this transfer more robust, we experimented with training additional models to serve as in-domain and out-of-domain classifiers. Although our results in this direction were inconclusive, we believe there is a promising path in newsworthiness detection (Spangher et al., 2023; Diakopoulos et al., 2010).

Our results have to be taken with some important caveats. While most newsworthiness models were impressively well-performing, many were not. Further exploration is needed to determine the causes. Additionally, despite the presence of non-English homepages in our dataset, we only tested with U.S.-based websites. We look forward to continuing to expand this work to address these concerns.

8 Conclusion

In this work, we introduced NewsHomepages, a comprehensive dataset capturing editorial prioritization decisions through the lens of homepage layouts across thousands of news organizations. By applying weakly-supervised models, we demonstrated the ability to robustly infer editorial judgments through pairwise comparisons of article positioning, achieving over 0.8 F1 score across various cases. These results underscore the consistency and predictability of editorial hierarchies encoded in homepage layouts.

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

9 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.

9.1 Computational Budget

The computational resources required for this project were substantial. The model training involved 4x A40 GPUs for initial phases and 16x 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.

9.2 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.

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

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.

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.

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.

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

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

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.

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.

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.

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.

Louis Rosenfeld. 2002. *Information architecture for the world wide web*. O'Reilly Media.

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.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*.

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*.

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).

	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.

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'

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"[8]

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.

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

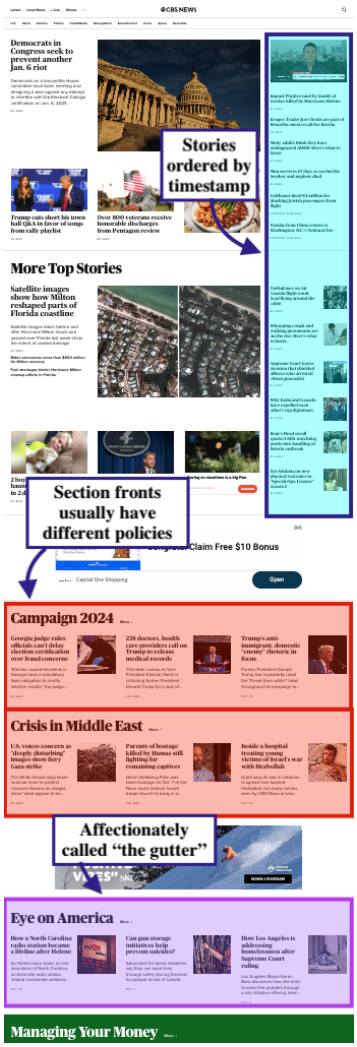
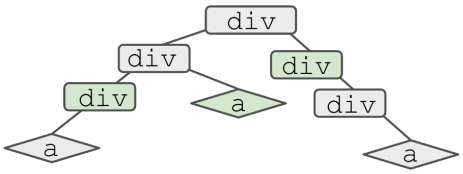


Figure 3: We show three sections of a sample homepage (from CBS News) where editorial decisions for different reasons. We highlight the "Breaking News" Section, "Section Fronts" and "The Footer".

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.

(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.

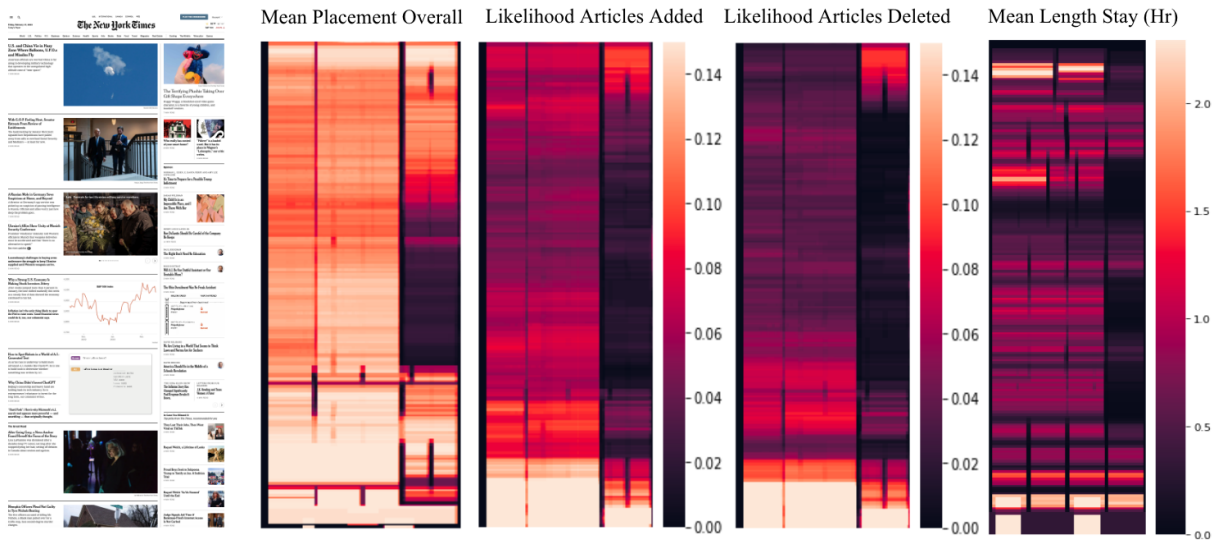


Figure 5: Different analyses we run on bounding boxes across time: average locations of bounding boxes on a homepage, locations where articles are added first, locations where they are removed, and the average time articles in various locations spend.

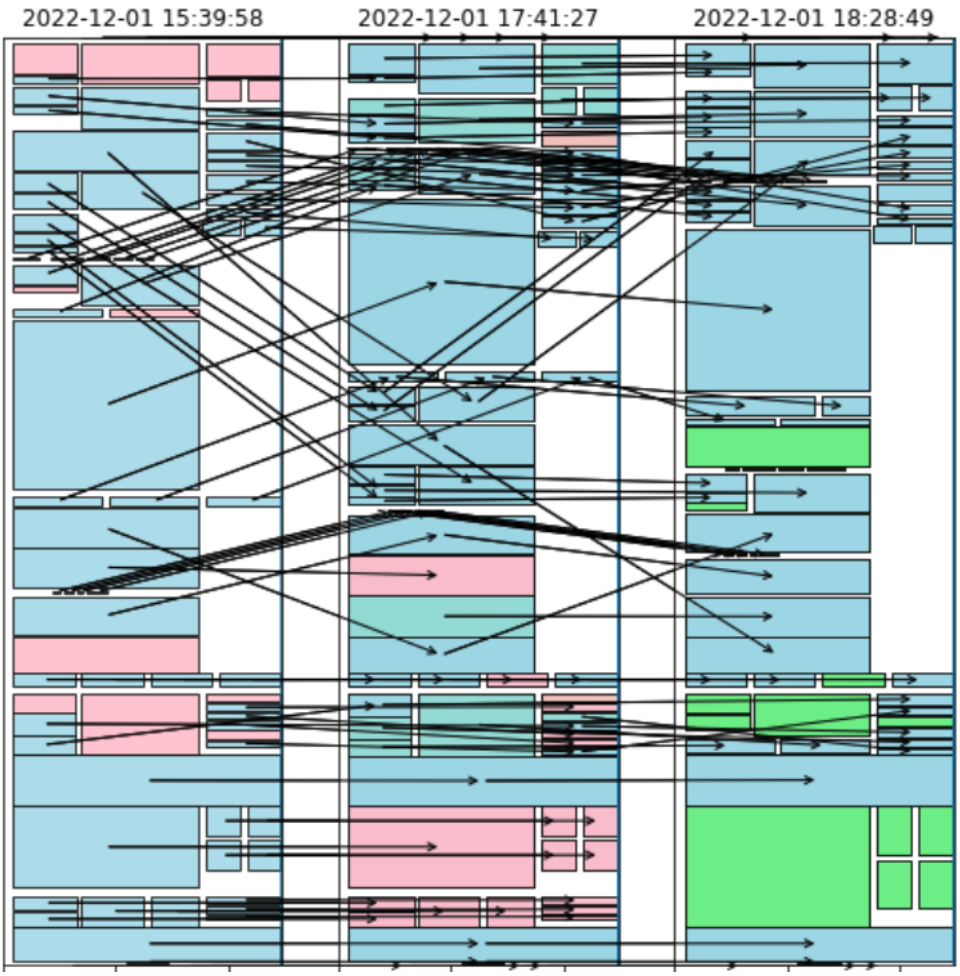


Figure 6: With our suite of tools for parsing homepages, we can examine on a granular level the movement of an article across the homepage.

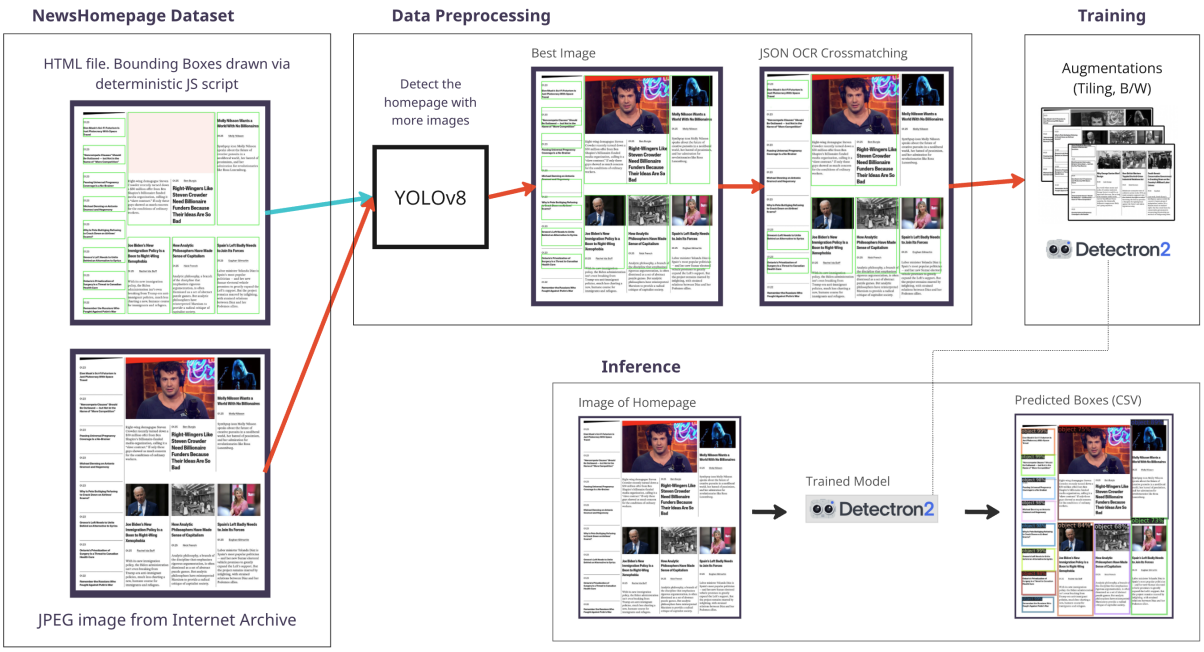


Figure 7: This diagram is an overview of the data preparation and training of the Detectron2 model for predicting bounding box on websites.

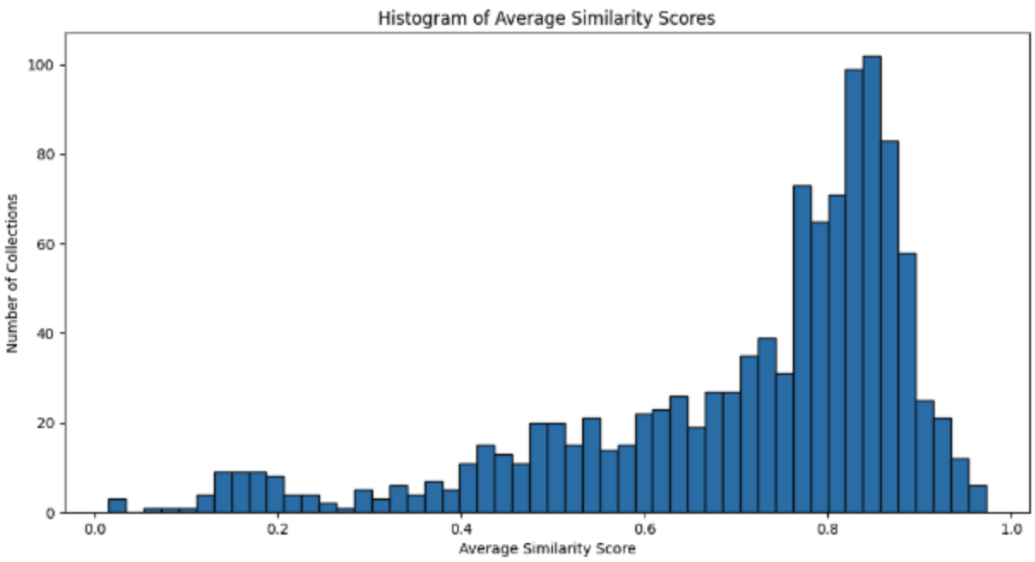


Figure 8: When sorting our sources to determine the ones most difficult for the DOM-Tree algorithm, we define the Average Similarity score to be a general measure as to how well the bounding box's text match the article's JSON file containing text/link pairs. High similarity score means high bounding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