# Analyzing Patterns and Influence of Advertising in Print Newspapers

**Anonymous submission**


## Abstract

This paper investigates advertising practices in print newspapers across India using a novel data-driven approach. We develop a pipeline employing image processing and OCR techniques to extract articles and advertisements from digital versions of print newspapers with high accuracy. Applying this methodology to five popular newspapers covering multiple regions and three languages (English, Hindi, and Telugu), we compile a rich dataset encompassing over 12,000 editions and hundreds of thousands of ads, representing an aggregated readership of over 100 million people. Using this extensive dataset, we conduct a comprehensive analysis to answer key questions about print advertising: who advertises, what they advertise, when they advertise, where they place their ads, and how they advertise. Our findings reveal significant patterns, including the consistent level of print advertising over the past six years despite declining print circulation, the overrepresentation of company ads on prominent pages, and the disproportionate revenue contributed by government ads. Furthermore, we examine whether advertising in a newspaper influences the coverage an advertiser receives. Through regression analyses on coverage volume and sentiment, we find strong evidence supporting this hypothesis for corporate advertisers. The results indicate a clear trend where increased advertising correlates with more favorable and extensive media coverage, a relationship that remains robust over time and across different levels of advertiser popularity.


## 1 Introduction

In this paper, we investigate advertising in *print* newspapers across India. Despite the exponential growth of digital news consumption, print newspapers continue to hold a significant position in the media landscape, both globally and within India.

Internationally, print media remains a vital source of information for many. According to a study by the Pew Research Center, 53% of Americans who prefer newspapers for local news opt for the print version over digital formats (Pew Research Center 2023a). The enduring trust in print media further emphasizes its importance. The Reuters Institute Digital News Report indicates that 62% of respondents trust print newspapers, compared to 44% for news websites (Reuters Institute 2023). This heightened level of trust is often attributed to the perceived editorial rigor and accountability inherent in print journalism. Economically, print media continues to be a substantial revenue generator for major news outlets. As of 2023, print circulation and advertising contributed approximately 60-70% of total revenue for several leading U.S. newspapers (Johnson and Smith 2023). The New York Times, for example, reported that about 30% of its total revenue was derived from print subscriptions and advertising (The New York Times Company 2023). Additionally, the effectiveness of print advertising bolsters its continued relevance. A study by NewsMediaWorks found that print newspaper advertisements were 1.5 times more effective in driving purchase intent compared to digital ads (NewsMediaWorks 2022).

In the Indian context, the significance of print media is even more pronounced. The Indian Readership Survey of 2020 (Council 2020), the primary source for newspaper readership data in India, reveals that India is the second-largest newspaper market globally, with over 110 million copies sold daily. Print media commands about 20% of India's total advertising expenditure, a stark contrast to the global average of just 4% (Farooqui 2024). The industry is poised for robust growth in 2025, driven by factors such as elections, with print advertising revenue anticipated to reach an all-time high.

Previous research analyzing advertising at scale has primarily focused on digital newspapers (Sridhar and Sriram 2015) or has been conducted on a small scale (Lindstädt and Budzinski 2011). This is largely due to the difficulty of obtaining large-scale data on print advertisements. Such data is either exclusively held by the newspapers themselves, making cross-paper comparisons impossible, or it is expensive to acquire from third-party sellers (Majid 2024), who may not even have data from countries like India.

In this paper, we obtain digital versions of print newspapers and use advances in image processing and Optical Character Recognition (OCR) techniques to develop a pipeline to extract articles and ads from print newspapers. We obtain 5 popular newspapers covering multiple regions and 3 languages from India (English, Hindi, and Telugu) with an aggregated reach of over 100 million people. We obtain articles spanning multiple years from these newspapers and apply our pipeline to extract with high confidence, *all* ads published in the paper along with their content.

Using this rich dataset, we answer, for the first time, *who* advertises, *what* they advertise, *when* (e.g. what time of the

week, time of year, etc), *where* (e.g. on what page) and how (e.g. display ads vs. text ads). We specifically focus on two important categories of advertisers – corporates and government, which together make up 45.55% of the total number of advertisements and account for about 47.32% of the total advertising expenditure. We reveal some interesting patterns in print advertising. For instance, despite the fact that print circulation is dropping significantly (Pew Research Center 2023b), print advertising remained consistent over the past 6 years. We uncover key patterns of advertisements by various companies and governments, including the over representation of company ads on prominent pages and the disproportionate revenue that government ads contribute to print advertising.

Moreover, using this data, we try to answer an important research question at scale – does advertising in a newspaper provide favorable coverage to the advertiser? We answer this question by regressing both the coverage volume and sentiment for various entities (companies and government). We find strong evidence for this hypothesis in the case of companies, indicating a clear trend in how advertising influences coverage as well as sentiment. We show that this relationship is robust to time and popularity.

The findings help us understand and analyze print media at scale for the first time providing a key source to complement analysis of digital media.

**Relevance to ICWSM**. While our study primarily focuses on print newspapers, its implications and methodologies are highly relevant to the web, particularly in the context of digital media and online advertising. The web has fundamentally transformed how news is produced, distributed, and consumed, with online platforms becoming the primary source of information for a vast majority of users. This shift has intensified the interplay between advertising and editorial content, making the investigation of media bias more pertinent than ever in the digital realm. The paper makes use of digital representations of print (or physical) newspapers, which are only accessible on the web. This allows us to do such a study at a scale which was not previously possible at this scale.

**Contributions**. In this paper, we make several key contributions to the study of print media advertising and its influence on news coverage:

- We introduce a novel pipeline capable of extracting and analyzing **all** advertisements from print newspapers. This pipeline is designed to handle large volumes of data efficiently, enabling extensive analysis of advertising content at an unprecedented scale.
- Applying our pipeline, we have compiled a unique and extensive dataset encompassing five different newspapers in India. This dataset covers four regions, three languages, and includes 12,358 editions, resulting in hundreds of thousands of advertisements. The diversity of the dataset offers a comprehensive representation of print media advertising across different linguistic and regional contexts.
- We provide a thorough examination of advertising strategies by addressing the fundamental questions of who advertises, what is advertised, how, where, and when. This multifaceted analysis sheds light on the behaviors and preferences of both government and corporate advertisers, revealing patterns in ad placement, timing, content, and size.
- We study the effect of advertising on media coverage by analyzing both the tone and volume of articles related to advertisers. Our findings demonstrate a clear relationship between advertising expenditure and the nature of coverage received, particularly for corporate entities. This contributes to the understanding of how advertising revenue may influence editorial content and potentially lead to media bias.
- Finally, we provide the code and datasets used in our approach to the community.[1] The code can help extend our analysis beyond the Indian context and the dataset can help further analysis of print news papers and their coverage.

In an era of increasing concerns about media bias, our dataset and analysis offer critical insights into the interplay between advertisers and media outlets. By enhancing transparency and understanding of advertising practices and their potential influence on news content, we contribute to the broader discourse on media integrity, informed citizenship, and the health of democratic processes. The methodologies and tools developed in our paper are not limited to the context of Indian newspapers but can be adapted to other regions and media formats, thus facilitating further research into the dynamics between advertising and media coverage globally.

## 2 Background and Related Work

**Media Bias and Advertising Influence**. The media is often regarded as the "Fourth Estate," underscoring its vital role in upholding democratic societies by providing unbiased and factual reporting. This concept emphasizes the media's responsibility to act as a watchdog, holding power structures accountable and fostering informed public discourse. According to McChesney (McChesney 2004), a healthy democracy necessitates a free and independent press that delivers diverse viewpoints without undue influence from external forces. However, maintaining editorial independence has become increasingly challenging due to commercial pressures (Siles and Boczkowski 2012) and the consolidation of media ownership. Bagdikian (Bagdikian 2004) highlights how media conglomerates can compromise journalistic integrity by prioritizing profit over the public interest, leading to homogenized content and potential biases. The tension between upholding democratic values and meeting commercial objectives creates a dilemma for media outlets. Hamilton (Hamilton 2004) explores how economic incentives influence news production, suggesting that market demands often shape what gets reported rather than purely journalistic considerations. In this context, auditing newspapers is crucial to ensure they adhere to ethical standards and continue to serve their role as the Fourth Estate.

The imperative for profitability significantly influences editorial decisions within media organizations. Picard

[1] http://bit.ly/4fZcIJw

(2011) examines the economics of media companies, highlighting that revenue generation—particularly through advertising—often takes precedence over journalistic ideals. This commercial focus can lead to content that appeals to advertisers or attracts larger audiences, potentially at the expense of investigative journalism or critical reporting. Croteau and Hoynes (2006) discuss how media outlets navigate the balance between maintaining editorial integrity and satisfying corporate interests. They argue that dependence on advertising revenue can result in self-censorship or the avoidance of topics that might alienate advertisers. This balancing act affects the type of news that gets published, as media companies strive to attract and retain advertisers while attempting to uphold the standards of independent journalism. Understanding these dynamics is essential for analyzing how economic factors shape media content and the implications for democratic discourse.

**Media Bias Favoring Advertisers**. Empirical studies have provided evidence that advertising revenue can influence editorial content, resulting in media bias favoring advertisers. Reuter and Zitzewitz (2006) examined the financial media and found that publications tend to offer more favorable coverage to firms that advertise with them. Their analysis revealed a positive correlation between the amount of advertising purchased by a company and the tone of editorial content it receives, suggesting that editors may compromise journalistic objectivity to maintain advertising relationships. Similarly, Di Tella and Franceschelli (2011) investigated the impact of government advertising on media coverage of corruption scandals in Argentina. They discovered that newspapers receiving substantial government advertising were less likely to report on corruption cases involving government officials. This dependence on advertising revenue from influential entities can lead to the suppression or under reporting of negative news about those entities. Similar trends have also been observed in India (Neyazi 2011; Dasgupta 2017).

Ownership structures and corporate interests significantly influence media content, often leading to biased reporting. Gentzkow and Shapiro (2010) analyzed U.S. daily newspapers and found that the slant of news coverage is affected by both the ideological preferences of the audience and the ownership of the media outlet. Their study suggests that media owners may cater to audience biases to maximize circulation and profits, resulting in a consistent content slant that aligns with specific viewpoints and potentially limits the diversity of perspectives available to the public. Gilens and Hertzman (2000) explored how corporate ownership affects news bias by examining newspaper coverage of the 1996 Telecommunications Act. They found that newspapers owned by companies with significant television interests provided more favorable coverage of the Act, which stood to benefit their corporate holdings. This indicates that corporate interests can directly influence editorial decisions, leading to reporting that serves the owners' financial agendas. Government entities and large corporations often use advertising spending as a tool to influence media coverage, potentially leading to favorable reporting or the suppression of negative news. Prat and Strömberg (2013) discussed the political economy of mass media, emphasizing how governments can manipulate news content through financial means such as advertising budgets. Their analysis highlighted that media outlets reliant on government advertising may avoid negative reporting on governmental actions to secure continued funding, thereby compromising journalistic independence.

**Challenges in print media analysis**. Obtaining large-scale data from print media poses significant challenges due to limited access, high costs, and the proprietary nature of data held by newspapers or third-party vendors. Gentzkow, Shapiro, and Sinkinson (2014) highlighted these difficulties in their historical analysis of U.S. newspapers, where data collection involved extensive archival research and manual digitization. The scarcity of digitized archives for print newspapers contrasts sharply with the abundance of data available from digital media, creating a barrier for researchers aiming to conduct comprehensive studies on print media content and its impact.

Puglisi and Snyder Jr (2015) discussed the obstacles in assessing media bias and content diversity due to these accessibility issues. The proprietary nature of print media archives often requires researchers to secure expensive licenses or subscriptions, limiting the scope and scale of potential studies. Peiser (2000) noted that declining newspaper readership has led to reduced investments in archiving and data preservation, further exacerbating the problem. Advancements in technology, particularly in Optical Character Recognition (OCR) and computational methods, have begun to mitigate some of the challenges associated with analyzing print media at scale. Smith (2007a) provided an overview of the Tesseract OCR engine, an open-source tool that has significantly improved the accuracy and efficiency of converting scanned print documents into machine-readable text. This development enables researchers to process large volumes of print media content, facilitating quantitative analyses that were previously impractical due to resource constraints.

In this work, we explore these dynamics in a fully automated manner by leveraging physical newspaper data to understand the ad ecosystem and quantify the impact of governmental and corporate advertisements on editorial bias. By employing computational techniques, we aim to analyze large-scale datasets, capturing the nuances of how advertisements affect the tone and volume of media coverage in Indian newspapers. To the best of our knowledge, we are the first to conduct such an analysis using an automated mechanism, and all data and code used in this study will be made publicly available for this purpose upon publication of the paper[2].

Our approach extends the existing literature by offering a data-driven, scalable method to assess the role of advertiser influence across various factors such as source and entity type.

## 3 Data

Our data collection involves scraping the archives of the print newspapers (also known as 'epapers') in India. We focus on five major news sources, which are among the coun-

[2]http://bit.ly/4fZcIJw

Table 1: Newspaper Data Comparison: The columns include the newspaper source, the time period of data collection, and the cities or zones covered. The number of editions represents the total editions (each daily issue of a newspaper specific to a city and source is considered an edition) analyzed, while the number of pages is the sum of pages in each edition.

| Source | Time Period | Cities / Zones | Num. Editions | Num. Pages | Num. Articles | Num. Ads |
|---|---|---|---|---|---|---|
| Hindustan Times | July 2019 - June 2024 | Delhi, Mumbai | 3,547 | 71,130 | 442,300 | 150,898 |
| Times Of India | Dec 2021 - Jun 2024 | Delhi, Mumbai, Chennai, Kolkata | 3,724 | 123,638 | 800,171 | 465,435 |
| Telegraph | May 2018 - Feb 2024 | Kolkata | 2,077 | 29,556 | 159,081 | 93,928 |
| Dainik Bhasker | Sept 2021 - Jun 2024 | Delhi | 1,016 | 13,995 | 56,031 | 14,862 |
| Sakshi | Oct 2022 - Jun 2024 | Hyderabad, Telangana, Andhra Pradesh | 1,994 | 32,465 | 162,232 | 111,250 |

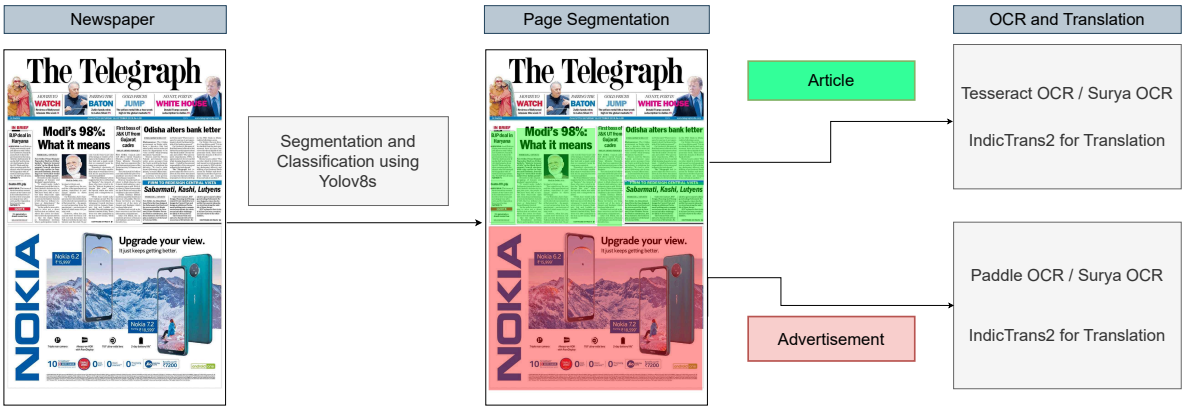



Figure 1: Processing epaper pages into textual entries across multiple sources and languages.

try's most popular: *The Times of India*, *Hindustan Times*, and *The Telegraph* (English sources), as well as *Dainik Bhaskar* (in Hindi), and *Sakshi* (in Telugu). Times of India is India's most circulated English newspaper; Hindustan Times is a close second and The Telegraph is also widely read and most circulated in the state of West Bengal. Dainik Bhaskar and Sakshi are among the most circulated Hindi and Telugu newspapers. For each source, we crawled epapers from the respective sites from the timelines mentioned in Table 1. We also collected papers from the same sources from multiple metropolitan cities to understand any regional effect.

The data collected primarily consists of page-level images. However, additional metadata was available for The Times of India and Hindustan Times. This metadata includes bounding box information for different page segments, division of articles and advertisements, and textual data corresponding to articles, though advertisements lacked associated text. This availability of metadata aided us in making a custom image segmentation model for processing other sources. Since the article text was already provided, we only needed to apply Optical Character Recognition (OCR) to the advertisements for these sources.

### 3.1 Image Segmentation Model

To analyze the distinction between advertisements and articles in newspaper pages, We utilized data from the Hindustan Times, which had information focusing on bounding boxes that indicate whether the content was an advertisement or an article. This data comprised 1,024 images, divided into 851 images for training, 151 images for testing, and 22 images for validation. After preprocessing the dataset with standard image preprocessing (fixing orientation and resizing), we finetuned YOLOv8s (Jocher, Chaurasia, and Qiu 2023) model for this task, which is particularly well-suited for object detection tasks like identifying object bounding boxes. The model's performance metrics are as follows: a mean Average Precision (mAP) of 96.8%, a Precision of 86.9%, and a Recall of 88.8%. These results demonstrate the model's high accuracy and robustness in distinguishing between advertisements and articles in our dataset and segmenting the epaper page properly into respective article/ad segments. Despite the availability of additional data, we observed diminishing returns when incorporating a larger dataset (around 10k samples) into the training process. Performance metrics for larger dataset can be found in Appendix A.

### 3.2 OCR and Translation

Once the Image Segmentation model processes the images, for each individual segment of the image identified as an advertisement or article, we apply a combination of OCR models based on the language and content type. For English-language articles, We use Tesseract OCR (Smith 2007b) because its page segmentation modes effectively retrieves text in the manner of reading order which makes it suitable for structured article content where text flow is crucial, moreover, it achieves an error rate of less than 5%, ensuring high accuracy in text retrieval.

However, Tesseract OCR faces significant challenges in English-language advertisements due to irregular layouts, diverse fonts, and random text arrangements, which reduce Tesseract's performance. To overcome this, we use PaddleOCR(PaddlePaddle Contributors 2024), which maximizes text extraction from more complex layouts, such as

advertisements, when the reading order is not a priority. Since the reading sequence in advertisements is generally less structured, PaddleOCR's robust extraction capabilities make it the ideal tool for this task.

For content in Indic languages (Hindi and Telugu in our dataset), we employ Surya OCR (Vik Paruchuri 2024a), as both PaddleOCR and TesseractOCR perform poorly with these languages and Surya OCR's performance being close to 1% error rate. Although Surya does not extract text in reading order, we address this limitation by leveraging its ordering and segmentation capabilities. We first use Surya to segment the text components in reading order, generate a new image based on these segments, and then apply the OCR process. Appendix C shows a detailed example of this process. Although Surya OCR does not perform as well on advertisement data compared to articles, it remains the most effective solution currently available for processing Indic-language text (Vik Paruchuri 2024b).

After extracting the Indic language text using Surya OCR, we translate it into English using the IndicTrans2 (Gala et al. 2023) model, which is capable of translating both Hindi and Telugu, ensuring that we can convert the regional language content into English for further analysis. Since our study does not primarily analyze the content of the articles and only uses keywords to filter different advertisers/content, the minor errors introduced by translation do not affect the study.

## 4 Analysis of the Print Ad Ecosystem

Our dataset offers a unique opportunity to analyze the nature of advertisements in print newspapers at an unprecedented scale. The methodology we have developed is highly scalable and applicable to any newspaper, regardless of language or nation, potentially providing much-needed transparency in the print advertising landscape.

We focus on differentiating between government and corporate advertising expenditures.[3] These two categories of advertisers roughly make up 50% of both the volume and spending on advertising. Advertising on the first, third, and last pages of a physical newspaper holds significant importance due to higher visibility and reader engagement (MediaScience 2024; Reynolds Journalism Institute 2024). These pages are considered premium spots and command higher advertising prices.

**Keyword Identification**. For the government analysis, we first identify articles related to government corruption using a manually curated set of keywords such as "bribe," "scam," "corrupt," and other relevant terms commonly found in corruption-related articles. Government advertisements are classified by detecting keywords like "government," "state," and "tender" in the ad text. The complete set of keywords used to identify both government articles and ads is provided in Appendix G. For the corporate analysis, we select a subset of the most prominent advertisers reported in TAM Media Research India's quarterly reports.[4] These companies represent key sectors such as Fast-Moving Consumer Goods (FMCG), Automobile, Technology, Education, and other major industries. Keywords associated with each company are derived from popular products and common terms used to refer to that company. A comprehensive list of the companies and their corresponding keywords can be found in Appendix H. Using these keywords, we extract relevant articles and advertisements for each entity, enabling a focused analysis of their advertising patterns.

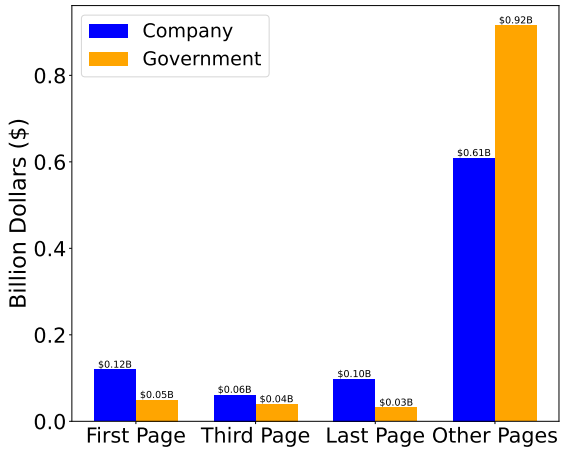



Figure 2: Spending by Companies and Government.

### 4.1 Who is advertising?

Figure 2 illustrates the absolute spending by government and companies. The data indicates that government entities are significant contributors to print advertising, surpassing corporate spending in total expenditure. Companies spend $890 million on ads in our dataset, whereas the government spends $1.02 billion. A majority of the ads in the newspapers are placed by the government.

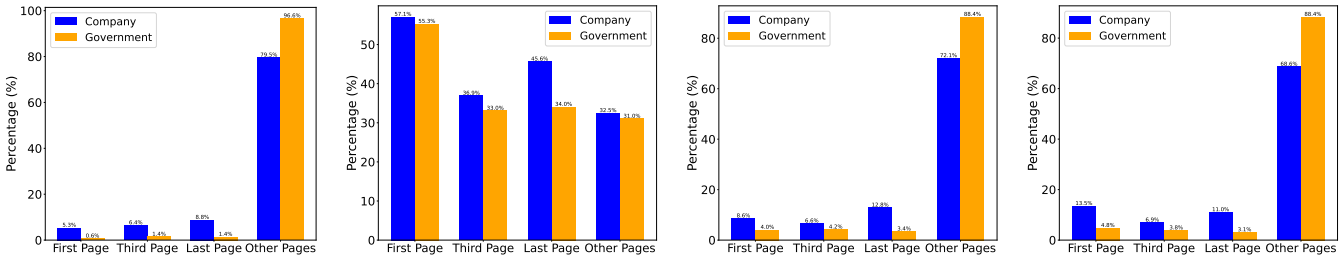
### 4.2 Where are the ads being placed?

Following industry conventions (detailed in Appendix Section F), we categorize ads based on their placement: **Premium Pages:** 1st, 3rd, and last pages. **Other Pages:** All remaining pages. To quantify the advertising presence and expenditure, we compute four main variables for each entity (government or companies):

1. The percentage of all ads provided by the entity that appear on a specific page of the newspaper.
2. The average percentage of the page area covered by the entity's ads when they appear on the specified page.
3. The percentage of the total specified page area occupied by the entity's ads, calculated across all ads by the entity. This provides an overall measure of the entity's visibility on the specified page, considering both the frequency and size of ad appearances.
4. The entity's total spending on newspaper ads, quantifying the entity's financial investment in their print advertising efforts.

[3]Unless explicitly specified, all the plots and analyses refer to the combined dataset; separate plots are available in the Appendix

[4]https://tamindia.com/




(a) Percentage of ads on a specific page

(b) Percentage of page area occupied by ads on average

(c) Percentage of page area in comparison to total area

(d) Percentage of ad expenditure compared to total ad spending

Figure 3: Where are ads being placed?

Our analysis yields the following results, as shown in Figures 3(a–d):

(i) **Ad Placement Frequency (Figure 3a):** A majority of government ads are not on the first, third, or last pages. Specifically, 88.4% of government advertising is found on other pages. Conversely, companies have a significantly higher fraction of ads on these premium pages, with 27.9% of their ad area appearing there.

(ii) **Ad Size on Premium Pages (Figure 3b):** When ads appear on the front page, government ads occupy an average of 55.3% of the page area, while company ads occupy 57.1%. This indicates that although the government advertises less frequently on premium pages, the size of their ads is comparable to that of companies when they do.

(iii) **Total Page Area Occupied (Figure 3c):** After normalizing by the total area of ads from either companies or the government, front-page ads contribute 8.56% of the total area for company advertisements. The combined 1st, 3rd, and last pages account for 27.9% of the company ad area, highlighting their emphasis on premium placements. In contrast, only 11.6% of government advertising is found on these pages.

(iv) **Advertising Expenditure Distribution (Figure 3d):** By calculating the sum of ad costs for specific pages and dividing by the total ad expenditure for company ads, we find that companies allocate 31.6% of their total ad expenditure to just the 1st, 3rd, and last pages. In contrast, government spending on advertisements is more evenly distributed across all pages, without disproportionate emphasis on specific pages.

## 4.3 How do they advertise?

To understand the methods and strategies employed by different advertisers, we analyze the sizes and placements of advertisements, which reveal insights into the types of ads provided, the approximate amounts spent, and the fraction of the page area they cover. By computing the size of the ads, we can discern patterns in advertising behavior between government entities and companies.

Figure 4 presents the cumulative distribution function (CDF) of advertisement sizes for both government and corporate advertisers. The data reveals a stark contrast between the two groups. Approximately 85% of the ads placed by government entities are small, occupying less than 10% of a page. Full-page ads by the government are exceedingly rare,

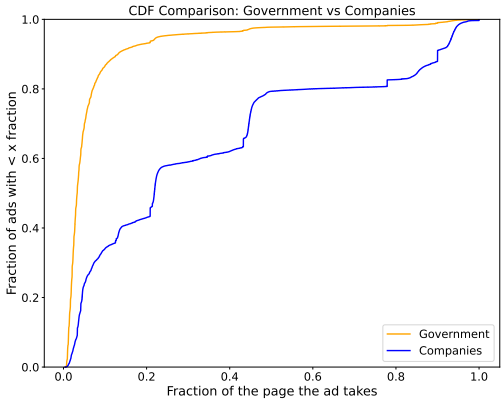



Figure 4: CDF of the area fraction of the ads showing government advertisers' strong preference for smaller ads (under 10%) and corporate advertisers' distinct focus on quarter-page, half-page, and full-page ads.

constituting only about 1% of their total advertisements. In contrast, corporate advertisers display a different pattern. The CDF for companies shows distinct jumps at the 25%, 50%, and 100% marks, corresponding to quarter-page, half-page, and full-page ads, respectively.

Further analysis uncovers additional observations regarding advertising strategies: We document an odd-page bias in advertisement placement. Advertisers prefer to place ads on odd-numbered pages, which are typically the right-hand pages in a newspaper layout. This preference may be an artifact of the way newspapers are read, as right-hand pages are more immediately visible when flipping through the pages. Our data supports this bias, with corporate ads appearing more frequently on odd pages (42,724 instances) compared to even pages (29,393 instances). This finding aligns with industry practices that consider odd-numbered pages as premium spots due to higher reader engagement.

Our dataset also allows us to compute the targeting priorities of various companies. For example, companies like Amazon and Patanjali have been primarily advertising in Hindi-language newspapers, as illustrated in Appendix Figure 12. We also observe interesting differences in advertising approaches across national English-language newspapers and regional publications. Companies tend to place more full-page ads in English-language newspapers, as shown in

Appendix Figure 10 . In general, most advertisements are heavily concentrated at smaller sizes, covering between 5% to 10% of a page, as depicted in Appendix Figure 9. Additionally, clear trends emerge when analyzing advertisement sizes across different industries and sectors. Educational institutions predominantly use full-page ads, as shown in Appendix Figure 13 . In contrast, insurance companies typically opt for half-page ads (Appendix Figure 14), whereas technology companies exhibit clustering around quarter-page, half-page, and full-page ads (Appendix Figure 15).

### 4.4 When are the ads being placed?

To understand the temporal dynamics of advertising in print newspapers, we conducted an analysis of time-based trends, as depicted in Figure 5. This figure illustrates the monthly average percentage of newspaper space occupied by advertisements from May 2018 to May 2024. Overall, the data reveals that advertisements consistently occupy approximately 35% of the total newspaper area throughout the years.

However, significant fluctuations are observed during certain periods: In early 2020, coinciding with the onset of the COVID-19 pandemic, there is a marked decline in the percentage of newspaper space devoted to advertisements. This drop reaches its lowest point in April and May 2020, where advertising space plummets to around 10%. The decline can be attributed to the economic uncertainty and reduced consumer activity during nationwide lockdowns, prompting advertisers to cut back on spending. A similar but smaller drop can also be found in June and July 2021, which aligns with the second wave of COVID-19 in India. This period was characterized by severe health impacts and renewed economic disruptions, leading to a temporary reduction in advertising activity.

As the largest English daily, with a readership of over 15 million, Times of India (TOI) consistently has a higher percentage of its space occupied by advertisements, averaging around 40%. This indicates a strong preference among advertisers for TOI, likely due to its wide reach among English-speaking and urban audiences. In contrast, Dainik Bhaskar, a leading Hindi-language daily, exhibits an average advertising space of approximately 20%, nearly half that of TOI.

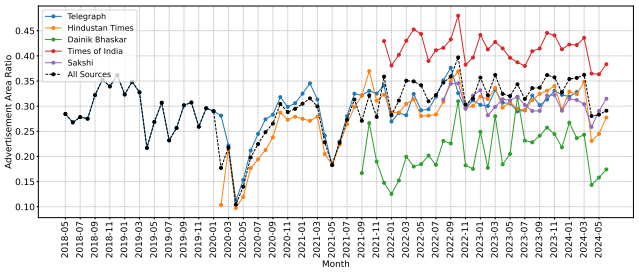



Figure 5: Monthly Advertisement Area Ratio by Source with Aggregate

Despite short-term fluctuations, the overall stability in the proportion of advertising space over the 6 year period suggests that print media remains a vital platform for advertisers. While digital media has grown substantially, print advertising continues to play a significant role, especially in reaching certain demographics and regions where newspapers are a primary source of information.

### 4.5 What are the ads about?

To understand the content focus of the advertisements, we conducted topic classification on the ad texts. We used an off-the-shelf topic classification model from Hugging Face,[5] specifically the RoBERTa-base model fine-tuned on a New York Times dataset.

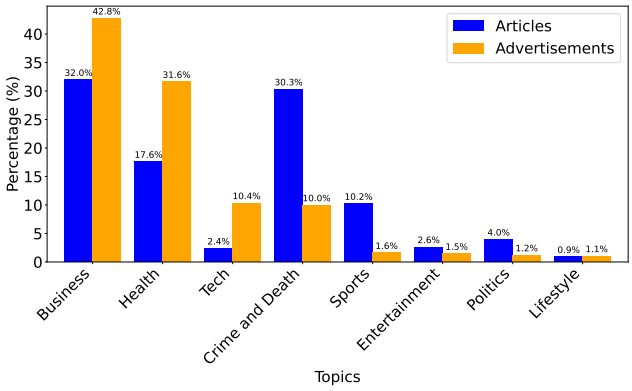



Figure 6: Topics covered in ads and in the article content.

This model provides classification into eight classes. Evaluation of the classification model can be found in Appendix B. Figure 6 presents the distribution of topics in the advertisements compared to the articles. Our analysis reveals several interesting observations. Business, health, and technology ads are significantly higher in proportion than articles in these topics. Conversely, topics such as crime and death, which are more prevalent in articles (including obituaries in the case of ads), are less represented in advertisements.

## 5 Do advertisers get positive coverage?

Building upon the data and insights presented in the previous sections, we address a pivotal research question: *Does advertising in a newspaper influence the coverage that advertisers receive in the news content?* Specifically, we aim to explore how advertising revenue from both corporate and government entities may affect the portrayal of news, regarding content bias and the extent of coverage. Leveraging our extensive dataset, we conduct a thorough analysis to examine this potential relationship.

Our investigation focuses on two primary dimensions: the **coverage** of entities, measured by the number of articles mentioning them, and the **sentiment** of that coverage, determined through sentiment of the articles. By examining these factors in relation to the advertising expenditure of the entities, we seek to understand whether there is a correlation

[5] https://huggingface.co/dstefa/roberta-base_topic_classification_nyt_news

between advertising spending and favorable media representation.

Inspired by previous literature that explores similar questions in specific contexts –such as the influence of car industry advertising on news coverage (Di Tella and Franceschelli 2011) or the reporting of government corruption (Smet and Vanormelingen 2012)– our approach builds upon established models from economics. We introduce measurable proxies for both advertising spending and media coverage to facilitate a quantitative analysis.

For advertising spending, we employ the *weighted ad ratio* as a proxy. This ratio quantifies the prominence and investment in advertisements by considering both the size of the ads and their placement within the newspaper. The ratio is calculated using the following formula:

$$\text{Weighted Ad Ratio} = \text{Scaling Factor} \cdot \frac{\text{Ad Area}}{\text{Page Area}}$$

The scaling factor is computed to indicate the relative importance and cost associated with the placement of an ad. Advertising rates vary depending on factors such as the page number, edition, and city. We obtained ad rates for various editions and page numbers for all the newspapers in our dataset to compute these scaling factors. For instance, if the base rate per square centimeter for an ad in The Times of India is $\$x$, and an ad on the front page costs $\$y$, the scaling factor for front-page ads is calculated as $y/x$. The scaling factors used in our analysis are based on actual rates from the newspapers, as detailed in Table 9. A comprehensive description of the costs and scaling factors can be found in Appendix F.[6]

By incorporating the scaling factor and normalizing by the page area, the weighted ad ratio allows for meaningful comparisons across different pages and publications, despite differences in page sizes and advertising costs. This normalization ensures that our analysis accounts for the relative prominence and investment in advertisements, rather than absolute measures that may be skewed by publication-specific characteristics.

Our regression analysis examines the relationship between the sentiment and volume of media coverage over a given time period and the weighted ad ratio of advertisements placed by the entities. By focusing on the difference in sentiment and coverage volume, we aim to explore whether increased advertising expenditure correlates with more favorable or extensive media portrayal, without relying on direct financial data that may be proprietary or unavailable.

Sentiment analysis of the articles was conducted using the model from (Camacho-Collados et al. 2022), one of the most widely adopted and downloaded models on Hugging Face providing sentiment classification of the content: -1, 0, or 1 corresponding to negative, neutral, or positive sentiment. Since we saw in Section 4 that government spending and types of ads are very different from other corporates, we will do the analysis separately.

[6]While we could directly use the estimated costs from Section B, variations in costs across advertisers and the approximate nature of these estimates led us to use the area of the ads, adjusted by scaling factors, as a more robust measure.

## 5.1 Regression setup

We employ a panel regression approach to understand the relationship between advertisement and news coverage, given that our dataset is structured with observations across multiple entities (such as news outlets and regions) over time.

The relationship between the advertising space and article sentiment (or coverage) is then analyzed using a panel regression model. Specifically, we regress the total sentiment of the articles (or coverage) on the weighted ad ratio occupied by the company's advertisements while incorporating fixed effects for media-company combinations and time (Beattie et al. 2021). The regression follows the structure:

$$\begin{aligned}\text{Total Sentiment}_{mnt} = \alpha + \beta \cdot \text{Weighted Ad Ratio}_{mnt} \\ + \gamma_{mn} + \delta_t + \epsilon_{mnt}\end{aligned}$$

where $\alpha$ represents the intercept, serving as the baseline level of the dependent variable. The coefficient $\beta$ quantifies the effect of the percentage of page area occupied by company $n$'s advertisements in newspaper $m$ during time period $t$, denoted as $\text{area}_{mnt}$. The term $\gamma_{mn}$ captures the media-company fixed effects, accounting for unobserved heterogeneity between different newspapers and companies. Time-fixed effects are incorporated through $\delta_t$, which accounts for variations over daily, weekly, or monthly time periods. Finally, $\epsilon_{mnt}$ is the error term that captures the unexplained variation in sentiment.

For companies, the $n$ refers to each company. For government, we consider all government ads as a single entity. We perform two types of regressions – one for sentiment and one for coverage. This formulation allows us to isolate the specific relationship between the percentage of advertisement area and the sentiment of articles while controlling for both unobserved differences across media-company pairs and time-based fluctuations and capture the relationship between advertisement percentage and sentiment. To also understand if popularity plays any role in sentiment and count of companies involved in the regression. We also consider a regression with an added popularity term (likely influencing how positively the coverage is framed) and with fixed effects on company and time. To obtain the popularity of a company, we used Google trends data and computed trend values for popularity for all the terms corresponding to a company.

## 5.2 Findings

For government, across most models, Tables 2 and 3, there is a negative relationship between weighted ad ratio and sentiment, indicating that an increase in the percentage of page area occupied by advertisements is associated with more negative sentiment in articles. This may be because, unlike companies that can selectively choose which newspapers to advertise in, government-related works (e.g. tenders and contracts) and announcements must be published in all major newspapers by law. This requirement removes the incentive for newspapers to tailor positive coverage in exchange

Table 2: Panel Regression Results: The Impact of Ad Page Percentage on Total Sentiment for Government Articles and Advertisements. Longitudinal count represents the number of entities, while the time period count represents the number of observations over time.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Dependent Variable: Total Sentiment** | | | | |
| **Total Ad Page Percent (Coefficient)** | -0.0730*** | -0.0115*** | -0.0709*** | -0.0084* |
| | (0.0047) | (0.0026) | (0.0070) | (0.0044) |
| **Fixed Effects: Source** | No | Yes | No | Yes |
| **Fixed Effects: Time** | No | No | Yes | Yes |
| **Entity Count** | 5 | 5 | 5 | 5 |
| **Time Period Count** | 75 | 75 | 75 | 75 |
| **$R^2$** | 0.9192 | 0.0385 | 0.8999 | 0.0179 |

Table 3: Panel Regression Results: The Impact of Ad Page Percentage on Count of Articles for Government Articles and Advertisements. Standard errors are clustered by entity.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Dependent Variable: Total Sentiment** | | | | |
| **Total Ad Page Percent (Coefficient)** | 0.5810*** | 0.1041* | 0.5337*** | 0.0214 |
| | (0.0078) | (0.0541) | (0.0093) | (0.0238) |
| **Fixed Effects: Source** | No | Yes | No | Yes |
| **Fixed Effects: Time** | No | No | Yes | Yes |
| **Entity Count** | 5 | 5 | 5 | 5 |
| **Time Period Count** | 75 | 75 | 75 | 75 |
| **$R^2$** | 0.9366 | 0.0618 | 0.9346 | 0.0032 |

for government advertisements, resulting in a more neutral or negative tone.

For companies, Tables 4, 5 demonstrates a robust and positive relationship between $\mathrm{Weighted\ Ad\ Ratio}$ and both $\mathrm{Sentiment}$ and total $\mathrm{Coverage}$. In all model specifications, the coefficient for ad page percentage remains statistically significant, indicating that increased advertising is associated with more favorable sentiment (and coverage) in news articles. Even when controlling for newspaper × company and time-fixed effects, the positive impact of advertising persists.

The results indicate strong and significant effects, showing that a 1% increase in a company's weighted ad ratio leads to an average increase of 0.0189 units in the total sentiment score. Given the narrow sentiment scale, which ranges from -1 to 1, this increase of 0.0189 units is substantial.

While considering the popularity of company terms from Google trends, We find a surprising relation, In Appendix Table 7, we notice that including the popularity term into the regression, the relationship between sentiment and advertising decreases and is non-significant when accounting for company fixed effects. This suggests that the influence of ad spending on sentiment may be less straightforward, where popularity might play a more complex role. However, overall, neither ad spending nor popularity shows consistent significance on sentiment across all specifications when company and time effects are included.

Even after accounting for popularity, weighted ad ratio remains highly significant across all models, indicating a robust relationship between ad page percentage and the count of articles (Appendix Table 8). This mirrors the results from Table 5, where ad page percentage alone was a strong predictor of article volume, suggesting that higher ad spending continues to drive media coverage, even when company popularity is taken into account. Notably, the effect of popularity itself is inconsistent across models, with a negative relationship in some models and non-significant results in others, highlighting that ad spending is a more reliable driver of media attention.

## 6 Conclusion

In this paper, we introduced a unique pipeline for collecting and analyzing advertisements in print newspapers at scale. Our methodology is highly scalable and applicable to any newspaper, irrespective of language or region, opening new avenues for large-scale research on the effects of advertisements in print media. By open-sourcing our code and dataset, we aim to enable researchers and practitioners to further explore the intricacies of print advertising and its impact on media content.

Our comprehensive analysis across multiple languages and newspapers revealed significant insights into the scale and nature of both government and corporate advertising. By examining who advertises, where, when, how, and what they advertise, we uncovered distinct patterns in advertising strategies. Government entities are major advertisers, with a wide distribution of smaller ads across various pages, while companies tend to invest in premium placements and larger ad formats to maximize visibility.

Through our regression analysis, we found a clear, statistically significant relationship between corporate advertis-

Table 4: Panel Regression Results: The Impact of Ad Page Percentage on Total Sentiment for Company Articles and Advertisements. Standard errors are clustered by entity.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Dependent Variable: Total Sentiment** | | | | |
| **Total Ad Page Percent (Coefficient)** | 0.0189*** | 0.0139*** | 0.0170*** | 0.0137*** |
| | (0.0035) | (0.0035) | (0.0027) | (0.0037) |
| **Fixed Effects: Newspaper × Company** | No | Yes | No | Yes |
| **Fixed Effects: Time** | No | No | Yes | Yes |
| **Entity Count** | 155 | 155 | 155 | 155 |
| **Time Period Count** | 72 | 72 | 72 | 72 |
| **$R^2$** | 0.1817 | 0.0268 | 0.1382 | 0.0244 |

Table 5: Panel Regression Results: The Impact of Ad Page Percentage on Count of Articles. Standard errors are clustered by entity.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Dependent Variable: Count Of Articles** | | | | |
| **Total Ad Page Percent (Coefficient)** | 0.4360*** | 0.2225*** | 0.4366*** | 0.2232*** |
| | (0.0265) | (0.0192) | (0.0280) | (0.0201) |
| **Fixed Effects: Newspaper × Company** | No | Yes | No | Yes |
| **Fixed Effects: Time** | No | No | Yes | Yes |
| **Entity Count** | 149 | 149 | 149 | 149 |
| **Time Period Count** | 70 | 70 | 70 | 70 |
| **$R^2$** | 0.8277 | 0.3606 | 0.8101 | 0.3546 |

ing expenditure and both the volume and tone of the media coverage they receive. This suggests that higher advertising spending by companies may correlate with more favorable and extensive coverage in the news content. In contrast, the relationship for government entities was less pronounced, possibly due to the obligatory nature of many government ads, which are legally mandated and broadly disseminated, making it challenging to influence coverage at scale. These findings raise important questions about media bias in favor of advertisers and its potential effects on public discourse, informed citizenship, and democratic processes. The influence of advertising revenue on editorial content underscores the need for transparency and ethical considerations within the media industry.

Our dataset remains highly underutilized, offering rich opportunities for further research. Notably, we possess the actual text of the articles that appeared adjacent to the advertisements, enabling studies on contextual influence and placement effects. Additionally, our dataset serves as a valuable resource for comparative analyses with digital media, allowing for explorations of cross-media advertising strategies and their implications.

## References


Bagdikian, B. 2004. H., 2004, The New Media Monopoly.

Beattie, G.; Durante, R.; Knight, B.; and Sen, A. 2021. Advertising Spending and Media Bias: Evidence from News Coverage of Car Safety Recalls. *Manage. Sci.*, 67(2): 698–719.

Camacho-Collados, J.; Rezaee, K.; Riahi, T.; Ushio, A.; Loureiro, D.; Antypas, D.; Boisson, J.; Espinosa-Anke, L.; Liu, F.; Martínez-Cámara, E.; et al. 2022. TweetNLP: Cutting-edge natural language processing for social media.

Council, M. R. U. 2020. Indian Readership Survey 2019 Q4. https://mruc.net/uploads/posts/cd072cdc13d2fe48ac660374d0c22a5d.pdf.

Croteau, D.; and Hoynes, W. 2006. *The business of media: Corporate media and the public interest*. Pine forge press.

Dasgupta, B. 2017. Tackling 'bias' and fake coverage in the Indian media.

Di Tella, R.; and Franceschelli, I. 2011. Government Advertising and Media Coverage of Corruption Scandals. *American Economic Journal: Applied Economics*, 3(4): 119–51.

Farooqui, J. 2024. Strong outlook: Indian print media industry is poised to record robust growth in 2024. https://economictimes.indiatimes.com/industry/media/entertainment/media/strong-outlook-indian-print-media-industry-is-poised-to-record-robust-growth-in-2024/articleshow/107801911.cms. [Accessed 13-10-2024].

Gala, J.; Chitale, P. A.; Raghavan, A. K.; Gumma, V.; Doddapaneni, S.; M, A. K.; Nawale, J. A.; Sujatha, A.; Puduppully, R.; Raghavan, V.; Kumar, P.; Khapra, M. M.; Dabre, R.; and Kunchukuttan, A. 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *Transactions on Machine Learning Research*.

Gentzkow, M.; and Shapiro, J. M. 2010. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1): 35–71.

Gentzkow, M.; Shapiro, J. M.; and Sinkinson, M. 2014. Competition and ideological diversity: Historical evidence

from us newspapers. *American Economic Review*, 104(10): 3073–3114.

Gilens, M.; and Hertzman, C. 2000. Corporate ownership and news bias: Newspaper coverage of the 1996 Telecommunications Act. *Journal of Politics*, 62(2): 369–386.

Hamilton, J. T. 2004. *All the news that's fit to sell: How the market transforms information into news*. Princeton University Press.

Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. Ultralytics YOLOv8.

Johnson, A.; and Smith, B. 2023. The State of News Media Finances. *Columbia Journalism Review*.

Lindstädt, N.; and Budzinski, O. 2011. Newspaper vs. Online Advertising–Is There a Niche for Newspapers in Modern Advertising Markets? *Online Advertising–Is There a Niche for Newspapers in Modern Advertising Markets*.

Majid, A. 2024. Top 25 US newspaper circulations: Print circulations of largest titles fall 14% in year to September 2023. https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/us-newspaper-circulation-2023/. Accessed: 13-10-2024.

McChesney, R. D. 2004. *The problem of the media: US communication politics in the twenty-first century*. NYU Press.

MediaScience. 2024. The Benchmark Series: The Powerful Impact of Ad Placement.

NewsMediaWorks. 2022. The Power of Print: Advertising Effectiveness in Newspapers. Research report, NewsMediaWorks.

Neyazi, T. A. 2011. India on television: how satellite news channels have changed the way we think and act.

PaddlePaddle Contributors. 2024. PaddleOCR: Awesome multilingual OCR toolkits based on PaddlePaddle (practical ultra lightweight OCR system, support 80+ languages recognition, provide data annotation and synthesis tools, support training and deployment among server, mobile, embedded, and IoT devices). https://github.com/PaddlePaddle/PaddleOCR. Accessed: 14-10-2024.

Peiser, W. 2000. Cohort replacement and the downward trend in newspaper readership. *Newspaper Research Journal*, 21(2): 11–22.

Pew Research Center. 2023a. Local News Preferences in the Digital Age.

Pew Research Center. 2023b. Newspaper Readership Demographics.

Picard, R. G. 2011. *The economics and financing of media companies*. Fordham Univ Press.

Prat, A.; and Strömberg, D. 2013. The political economy of mass media. *Advances in economics and econometrics*, 2: 135.

Puglisi, R.; and Snyder Jr, J. M. 2015. Empirical studies of media bias. In *Handbook of media economics*, volume 1, 647–667. Elsevier.

Reuter, J.; and Zitzewitz, E. 2006. Do Ads Influence Editors? Advertising and Bias in the Financial Media. *The Quarterly Journal of Economics*, 121(1): 197–227.

Reuters Institute. 2023. Digital News Report 2023.

Reynolds Journalism Institute. 2024. The importance of ads on article pages: Balancing revenue and reader experience.

Siles, I.; and Boczkowski, P. J. 2012. Making sense of the newspaper crisis: A critical assessment of existing research and an agenda for future work. *New Media & Society*, 14(8): 1375–1394.

Smet, D.; and Vanormelingen, S. 2012. The Advertiser is Mentioned Twice. Media Bias in Belgian Newspapers.

Smith, R. 2007a. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, 629–633. IEEE.

Smith, R. 2007b. An Overview of the Tesseract OCR Engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 629–633. Washington, DC, USA: IEEE Computer Society. ISBN 0-7695-2822-8.

Sridhar, S.; and Sriram, S. 2015. Is online newspaper advertising cannibalizing print advertising? *Quantitative Marketing and Economics*, 13: 283–318.

The New York Times Company. 2023. 2023 Annual Report.

Vik Paruchuri. 2024a. GitHub - VikParuchuri/surya: OCR, layout analysis, reading order, table recognition in 90+ languages. https://github.com/VikParuchuri/surya. [Accessed 14-10-2024].

Vik Paruchuri. 2024b. GitHub - VikParuchuri/surya: OCR, layout analysis, reading order, table recognition in 90+ languages. https://github.com/VikParuchuri/surya?tab=readme-ov-file\#limitations. [Accessed 14-10-2024].

## A Segmentation Model and Outputs

The image segmentation model processes newspapers' pages and identifies segments that are either articles or advertisements. One such example is given below


Figure 7: Newspaper page before and after segmentation

The segments are then cropped and processed through the rest of the pipeline. The segmentation model processes the images with confidence and IOU thresholds of 0.1, and then the identified regions are cropped with a margin of 10 pixels. The model and code used for inference will be made public.

We also used 1024 pages for training the model despite having access to larger training datasets, primarily since the performance degraded with larger datasets. Given the importance of a balanced approach in identifying both articles and advertisements accurately, we selected the first model for further analysis due to its superior mAP, which provides a better overall measure of detection performance across all classes.

Table 6: Results for ad detection based on training data size

| Dataset Size | mAP | Precision | Recall |
|---|---|---|---|
| 1024 | 96.8% | 86.9% | 88.8% |
| 9329 | 95.3% | 83.5% | 97.8% |

## B Exploratory analysis plots

Approximately 2.4 million articles and advertisements were identified across the analyzed news sources. A detailed breakdown of the number of articles and advertisements per source is provided in Table 1. To understand the distribution of topics covered in the articles and advertisements, we employed a topic classification model that was fine-tuned using a dataset from The New York Times[7] with an accuracy and F1 score of 0.91. This model allowed us to categorize the content into predefined topics, providing a clearer understanding of the thematic focus within articles.

[7]https://huggingface.co/dstefa/roberta-base_topic_classification_nyt_news

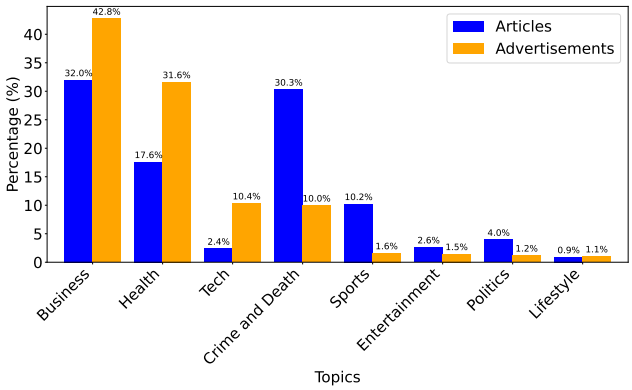



Figure 8: Distribution of Topics in Articles and Advertisements

Figure 9 shows the distribution of ad ratios across all papers. Most ads are small, occupying less than 10% of the space.

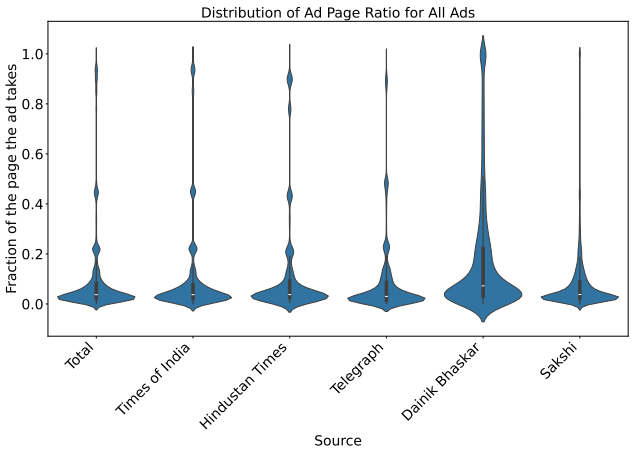



Figure 9: All ads across papers.

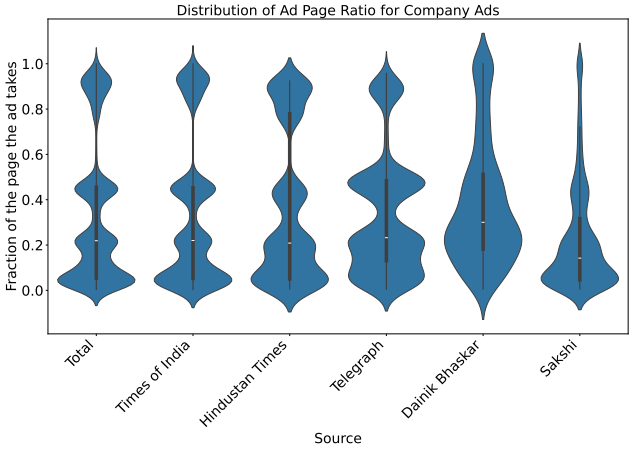



Figure 10: Company ads across papers

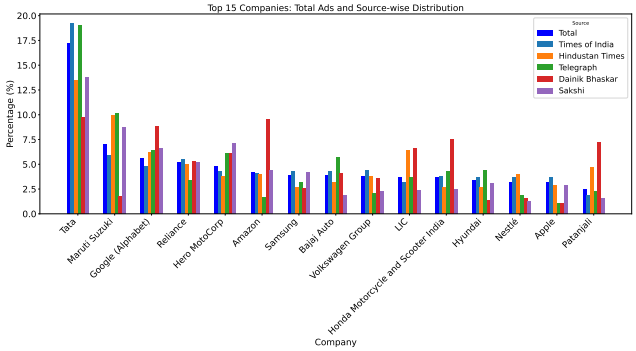



Figure 12: Percentage of ads provided by the top 15 advertisers in the 5 papers in our dataset.

Figure 10 shows the distribution of ad ratios across the 5 papers we study. Companies typically give larger ads. We can see that there is significant distribution around quarter, half, and full-page advertisements.

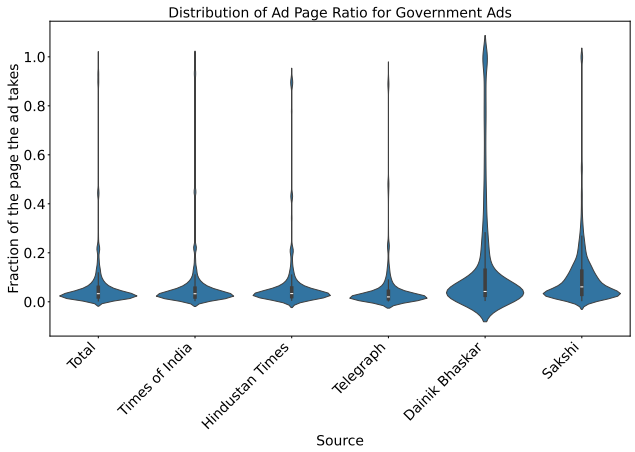



Figure 11: Government ads across papers

Figure 11 shows the ads by governments. We can see that across sources, the area page ratio is small, indicating that most ads are small in comparison to the page.

Figure 12 presents the distribution of the number of ads among the top 15 companies, broken down by contributions from various newspapers. Companies like Tata and Maruti Suzuki have a broad presence across multiple sources, while others focus their spending more selectively. This chart offers insight into the target audience for their industry.

Certain sectors specifically give certain types of ads. e.g., education sector typically gives larger full-page ads (figure 13. Insurance typically gives half-page ads (figure 14)

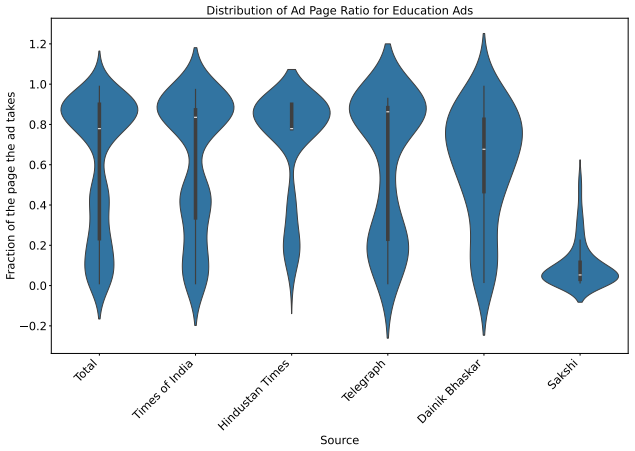



Figure 13: Education ads across papers - typically, education ads are mostly full page, except on Sakshi.

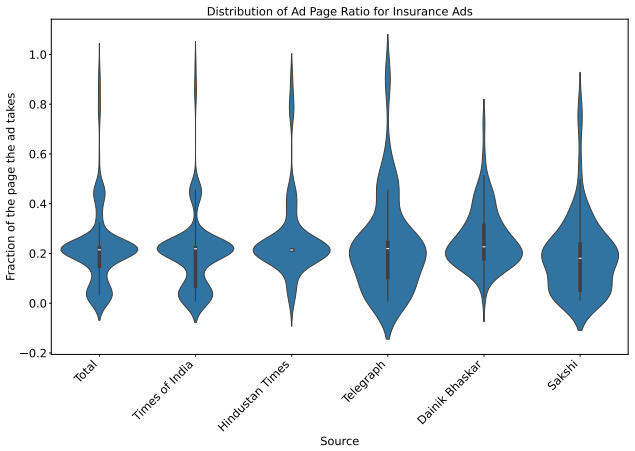



Figure 14: Insurance ads across papers - insurance ads are typically half-page

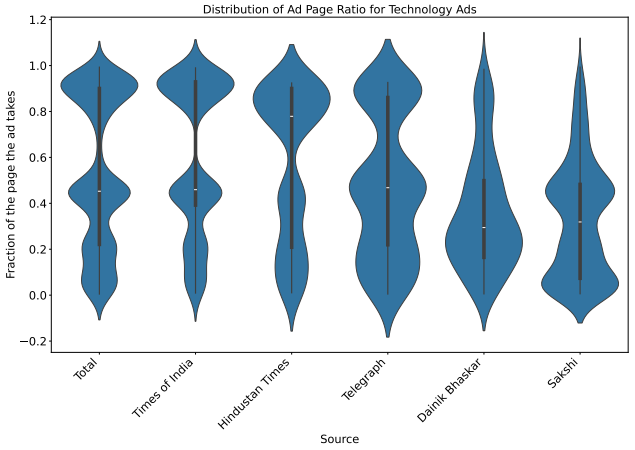



Figure 15: Technology ads across papers

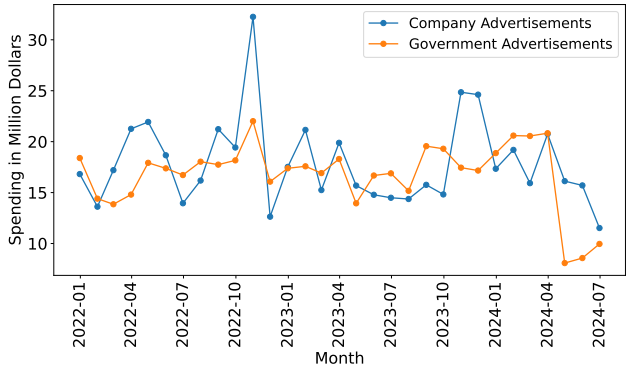



Figure 17: Spending by Companies and Government over Time.

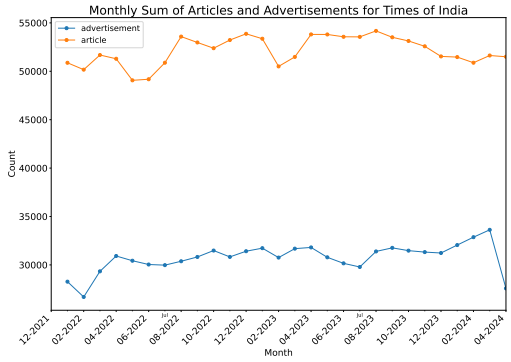



Figure 18: Count of Articles and Advertisements across time for Times of India

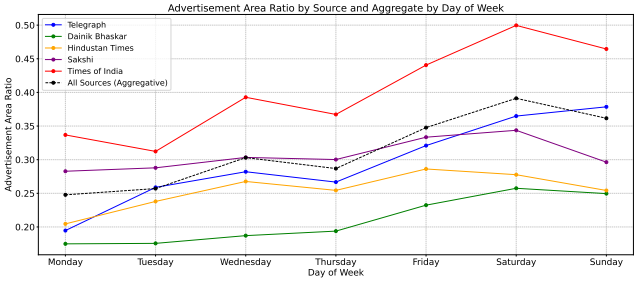



Figure 16: Advertiser Coverage by Day of the Week - Weekends show an increasing trend in Advertisement area

In Figure 17, we can see how both entities spend over time for a source, Times of India. Company spending exhibits clear seasonality, with fluctuations that may correspond to specific periods. Government spending appears more stable over time, with fewer pronounced peaks and stays close to company spending.

We can also observe the raw counts of articles and advertisements for various sources below.

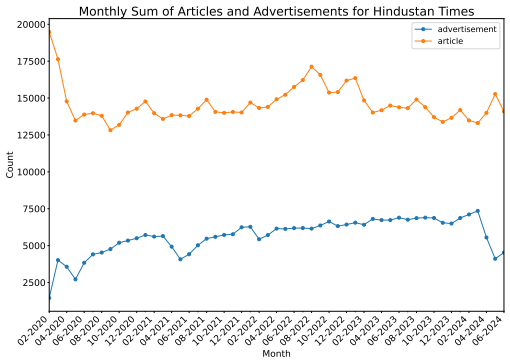



Figure 19: Count of Articles and Advertisements across time for Hindustan Times

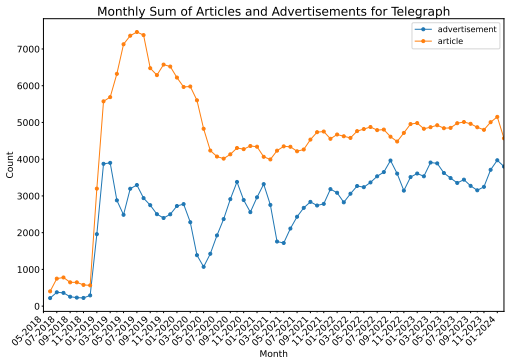



Figure 20: Count of Articles and Advertisements across time for Telegraph

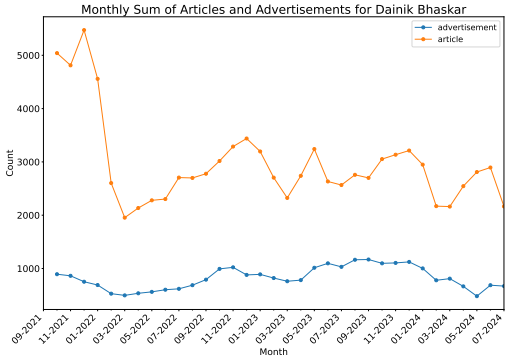



Figure 21: Count of Articles and Advertisements across time for Dainik Bhaskar

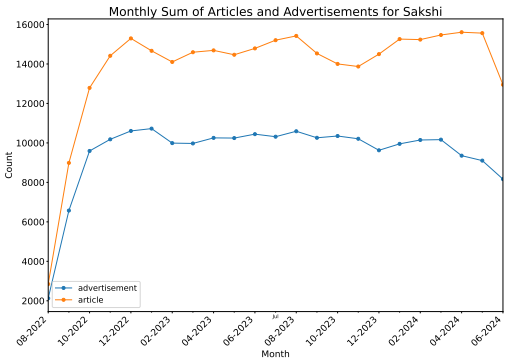



Figure 22: Count of Articles and Advertisements across time for Sakshi

## C Surya OCR in Reading Order

For English Articles, We leverage Tesseract OCR and its Page Segmentation Modes to extract text in reading order, but due to no similar functionality in Surya OCR, We first generate a reading order from the image segment using Surya OCR, then process the image into a manner which is sequentially in reading order and then run Surya OCR on the processed image. A detailed run of the process can be seen below

## D Impact of Advertising and Scandals on Sentiment

Figure 24 clearly demonstrates the influence of advertising on sentiment, particularly in how scandals impact conglomerates. This conglomerate was significantly affected by negative media coverage following a scandal in Jan 2023, prompting an increase in advertising efforts to mitigate the fallout. The data highlights the response, which is buying more ads, and gradually, the sentiment shifts following a scandal and gradually returns to pre-scandal levels.

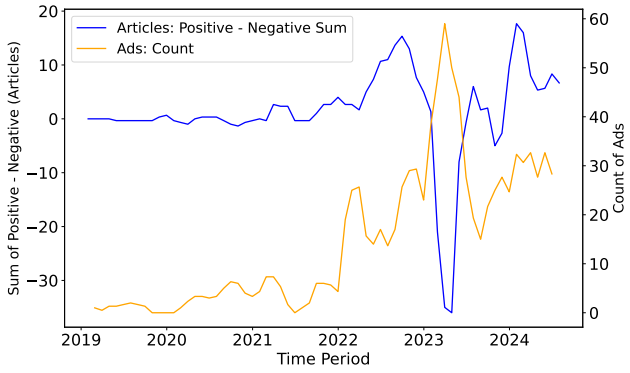



Figure 24: Influence of Advertising on Sentiment for Adani Conglomerate

## E Regression Tables

## F Price Variations and Scaling Factor

To better account for price variations across different pages and provide a more accurate measure of advertising expenditures, we apply a scaling factor to pages of particular interest, where advertising costs are typically higher. These pages include the first page, the third page, and the last page, all of which command premium rates. Additionally, we gather city-specific and source-specific advertising rates from online sources. To facilitate regression analysis and make the coefficients more interpretable, each entry is normalized by dividing by the minimum rate (546) in the dataset. This normalization allows for a clearer examination of the relationship between advertising expenditures and media coverage. The rates used for the table can be found here for the **Times of India**[8], **Hindustan Times**[9], **The Telegraph**[10], **Dainik Bhaskar**[11], and **Sakshi**[12].

## G Keywords To Identify Government-based articles and advertisements

The keywords used to identify Government-based articles and ads can be found in Table 10. Articles are identified if there is a match with any keyword from the corruption-related list and the presence of a government-related term (”government,” ”govt,” ”state,” or ”central”) to ensure government corruption-related articles. In contrast, advertisements are classified if they contain any keyword from the advertisement-related list.

## H Keywords To Identify Companies

The keywords are mostly curated by adding subsidiaries and brands that are uniquely identifiable to a brand. Table 11.

[8] https://web.archive.org/web/20241001131208/https://riyoadvertising.com/times-of-india-display-ad-rates.html

[9] https://web.archive.org/web/20241001131804/https://www.hindustantimes.com/rate-card/Impactht

[10] https://www.bhavesads.com/the-telegraph/display-ad-rates

[11] https://web.archive.org/web/20240503230534/http://www.riyoadvertising.com/dainik-bhaskar.html

[12] https://web.archive.org/web/20241001134917/https://riyoadvertising.com/sakshi.html

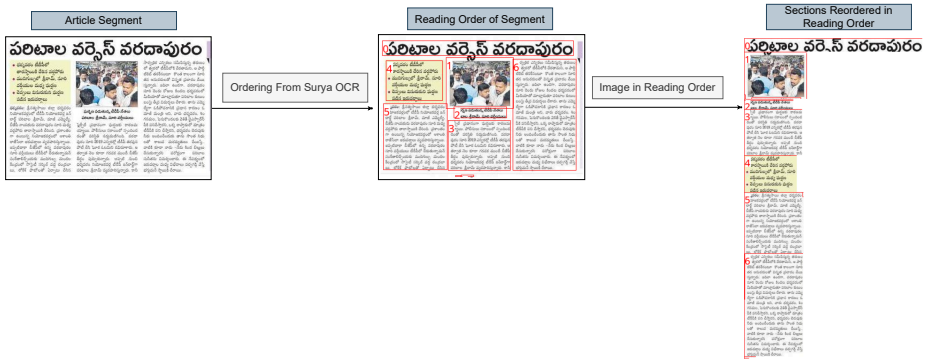



Figure 23: Processing Indic Image Segments

Table 7: Panel Regression Results: The Impact of Ad Page Percentage and Popularity on Total Sentiment. Standard errors are clustered by entity.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Dependent Variable: Total Sentiment** | | | | |
| **Total Ad Page Percent (Coefficient)** | -0.0003 | 0.0018 | -0.0016 | -0.0011 |
| | (0.0031) | (0.0015) | (0.0019) | (0.0014) |
| **Popularity (Coefficient)** | 0.0032*** | 0.0014 | 0.0041** | 0.0004 |
| | (0.0009) | (0.0017) | (0.0018) | (0.0013) |
| **Fixed Effects: Company** | No | Yes | No | Yes |
| **Fixed Effects: Time** | No | No | Yes | Yes |
| **Entity Count** | 40 | 40 | 40 | 40 |
| **Time Period Count** | 1372 | 1372 | 1372 | 1372 |
| **$R^2$** | 0.0625 | 0.0004 | 0.0110 | 7.399e-05 |

Table 8: Panel Regression Results: The Impact of Ad Page Percentage and Popularity on Count Of Articles. Standard errors are clustered by entity.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Dependent Variable: Count Of Articles** | | | | |
| **Total Ad Page Percent (Coefficient)** | 0.0640*** | 0.0185*** | 0.0299*** | 0.0214*** |
| | (0.0154) | (0.0042) | (0.0087) | (0.0078) |
| **Popularity (Coefficient)** | 0.0246*** | 0.0109 | -0.0163*** | 0.0136 |
| | (0.0041) | (0.0069) | (0.0061) | (0.0079) |
| **Fixed Effects: Company** | No | Yes | No | Yes |
| **Fixed Effects: Time** | No | No | Yes | Yes |
| **Entity Count** | 40 | 40 | 40 | 40 |
| **Time Period Count** | 1372 | 1372 | 1372 | 1372 |
| **$R^2$** | 0.5649 | 0.0118 | 0.0686 | 0.0150 |

Table 9: Page rates for various papers.

| Source | City | 1st Page | 3rd Page | Back Page | Base Price |
|---|---|---|---|---|---|
| Times of India | Mumbai | 9665 | 6850 | 7230 | 5640 |
| Times of India | Delhi | 6355 | 4830 | 5075 | 4120 |
| Times of India | Kolkata | 2435 | 1920 | 2105 | 1835 |
| Times of India | Chennai | 2815 | 2381 | 2405 | 1985 |
| Hindustan Times | Mumbai | 5100 | 3750 | 3750 | 3000 |
| Hindustan Times | Delhi | 10750 | 7470 | 7470 | 5970 |
| Dainik Bhaskar | Delhi | 867 | 661 | 774 | 546 |
| Sakshi | Andhra | 6739 | 2995 | 5990 | 2995 |
| Sakshi | Hyderabad | 2700 | 1200 | 2400 | 1200 |
| Sakshi | Telangana | 2700 | 1200 | 2400 | 1200 |
| Telegraph | Kolkata | 2641 | 2565 | 2430 | 2230 |

Table 10: Advertisement and Corruption Keywords.

| Category | Keywords |
|---|---|
| Advertisement Keywords | government, state, central, tender, gov, e-tender, corrigendum, e-corrigendum, govt., tenders, procurement, e-procurement |
| Corruption Keywords | bail, black, bribe, cbi, chor, conspiracy, corrupt, croni, demonet, expos, helicopt, investig, jail, jumla, lokpal, loot, nirav, probe, prosecut, rafal, raid, scam, scandal, steal, theft, thief, illegal, fraud, embezzle, misappropriat, laundering, offshore, tax evasion |

Table 11: Company keywords.

| **Company** | **Keywords** |
|---|---|
| Tata | Tata , Jaguar Land Rover , Taj Hotels , BigBasket , 1mg , AirAsia , Vistara , Tanishq , Titan , Starbucks , Voltas , Vivanta , Air India , Croma |
| Reliance | Reliance , JioFiber , JioMart , AJIO , Netmeds , Hamleys , Urban Ladder |
| Hindustan Unilever | Hindustan Unilever , HUL , Lakmé , Lifebuoy , Dove , Surf Excel , Kwality Wall's , Bru , Kissan , Vaseline , Ponds , Pepsodent , Clinic Plus , Rin , Axe |
| Procter & Gamble (P&G) | Procter & Gamble , Procter and Gamble , P&G , Pampers , Ariel , Tide , Gillette , Whisper , Vicks , Olay , Pantene , Head & Shoulders , Oral-B , Old Spice |
| ITC Limited | ITC Limited , Sunfeast , Aashirvaad , Savlon , Fiama , Vivel , ITC Hotels , Bingo! , Yippee! , Classmate , Wills , Gold Flake |
| Godrej Group | Godrej , Good Knight , Cinthol |
| Bharti Airtel | Airtel , Wynk Music |
| Samsung | Samsung |
| Xiaomi | Xiaomi , Redmi , POCO , Mi TV , Mi Smart Home , Mi Ecosystem , MIUI |
| Vivo | Vivo |
| Oppo | Realme , Oppo |
| OnePlus | OnePlus |
| Maruti Suzuki | Maruti Suzuki , Suzuki , Nexa |
| LIC | Life Insurance Corporation of India , LIC |
| Hyundai | Hyundai Motor India , Kia , Hyundai |
| Toyota Kirloskar | Toyota |
| Renault India | Renault , Dacia |
| MG Motor India | Morris Garages , MG Hector , MG Motor |
| Stellantis | Stellantis , Jeep India , Citroen India , Fiat , Mopar |
| BMW Group India | BMW |
| Mercedes-Benz India | Mercedes-Benz |
| Amazon | Amazon , Kindle |
| Coca-Cola | Thums Up , Sprite , Fanta , Minute Maid , Kinley , Maaza , Coca-Cola , Diet Coke , Smartwater |
| PepsiCo | PepsiCo , Pepsi , Mirinda , 7Up , Lay's , KurKure , Tropicana , Mountain Dew , Gatorade , Quaker Oats |
| Adani Group | Adani , Ambuja Cements |
| Mahindra Group | Mahindra & Mahindra , Mahindra Tractors , Tech Mahindra , Mahindra Finance , Mahindra Electric , Club Mahindra , Mahindra Lifespaces , Automobili Pininfarina |
| Nestlé | Nestlé India , Maggi , Nescafé , KitKat , Milo , Milkmaid , Nestea , Cerelac , Everyday , Perrier |
| Sony | Sony , PlayStation , SonyLIV |
| Volkswagen Group | Volkswagen , Audi , Porsche , Bentley , Lamborghini , Skoda , Bugatti , Ducati |
| Ford Motor Company | Ford |
| Apple | iPhone , iPad , MacBook , Apple Watch , iMac , Apple TV , Apple Music , Apple Pay , iCloud , Apple Store |
| Google (Alphabet) | Google Search , YouTube , Google Maps , Google Cloud , Google Ads , Android , Google Play , Gmail , Google Pixel , Nest |
| Hero MotoCorp | Hero MotoCorp , Splendor , HF Deluxe , Passion , Glamour , Xpulse , Hero MotoSports |
| Honda Motorcycle and Scooter India | Honda , Activa |
| Bajaj Auto | Bajaj , Pulsar , Dominar , Avenger , KTM , Husqvarna , Bajaj Finserv , Bajaj Finance |
| FIITJEE | FIITJEE |
| Byju's Aakash | Byju's , Aakash |
| Allen Career Institute | Allen |
| Nissan | Nissan , Datsun |
| Prestige | Prestige TTK |
| BigBasket | BigBasket , BB Daily , BBinstant |
| Amul | Amul |
| Patanjali | Patanjali |