Patterns and Influence of Advertising in Print Newspapers

N Harsha Vardhan, 2020111022
CSD , UG2k20.

Advised by Prof Ponnurangam Kumaraguru

Problem & Motivation

Print media remains highly trusted and widely read in India and globally. Media integrity impacts public discourse and democracy

Does advertising revenue compromise editorial independence?

Lack of large-scale empirical studies on print newspapers

Problem & Motivation

Print media remains highly trusted and widely read in India and globally.

Print media is public trusted, especially in India, with a readership of over **110 million** daily (IRS 2020).

Globally too, print outperforms digital on trust metrics: 62% trust print newspapers vs. 44% for news websites (Reuters Institute, 2023).

Problem & Motivation

Commercial pressures and media consolidation challenge this trust. (McChesney, 2004; Bagdikian, 2004; Siles, 2012)

Fragile balance between editorial independence and advertiser interests especially in a country like India where print ad revenues make a larger share of media income than the global average (Javed Farooqui, 2024)

‘Sarkari bulletin’: How the Indian state has used fear and ad money to hollow out Kashmir’s press

Five months into the lockdown, most editors are forced to function as censors and ‘gatekeepers’ of the government.

By: Rayan Naqash 05 Jan, 2020

Gaps in the Literature

Large-scale empirical work on print media advertising remains sparse due to data inaccessibility. (Gentzkow, 2014)

Prior studies have tested similar hypotheses, correlations between advertisers and media coverage—

1. Government ad influence in Argentina (Di Tella & Franceschelli, 2011)
2. Corporate ad influence in U.S. financial media (Reuter & Zitzewitz, 2006)
3. Bias from ownership and political alignment (Gentzkow & Shapiro, 2010; Gilens & Hertzman, 2000)

Gaps in the Literature

However these studies are extremely specific, with manually curated datasets with limited transparent and replicability.

Government ad influence in Argentina (Di Tella & Franceschelli, 2011) ➔ only looks at corruption and front page articles, about 1000 data points.

Corporate ad influence in U.S. financial media (Reuter & Zitzewitz, 2006) ➔ only looks at fund suggestions by newspaper columnists, about 10k data points.

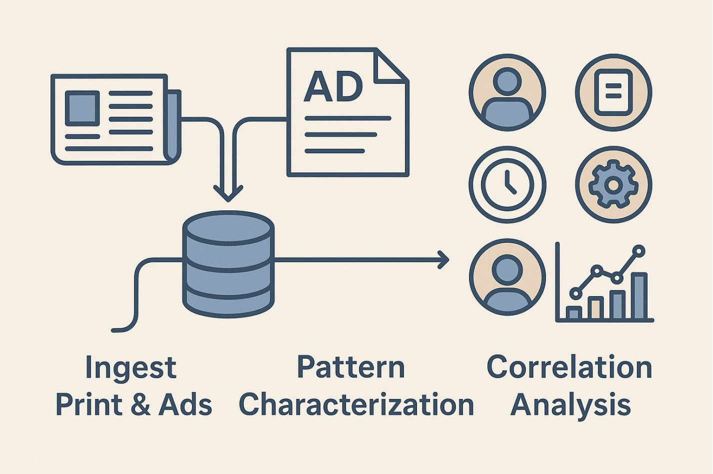
Bias from ownership and political alignment (Gilens & Hertzman, 2000) ➔ only looks at ownership structures and newspaper responses to Telecommunications Act 1996, 113 hand annotated articles.

Research Objectives

Develop a scalable, language-agnostic pipeline to analyze print content

Characterize advertising patterns by: Who, What, When, Where, How

Assess whether advertising expenditure correlates with editorial coverage



Data Collection

Five major Indian newspapers (English: TOI, HT, Telegraph; Hindi: Dainik Bhaskar; Telugu: Sakshi).

The sources were primarily chosen based on readership and availability of epapers over considerable time.

Over **12,000** editions; three languages; multiple cities. The final dataset has about **1.6M** articles and **830K** advertisements.

Hundreds of thousands of ads extracted via image scraping and metadata.

Data Collection

Table 3.1: Newspaper Data Comparison: The columns include the newspaper source, the time period of data collection, and the cities or zones covered.

Source	Time Period	Cities / Zones	# Editions	# Pages	# Articles	# 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

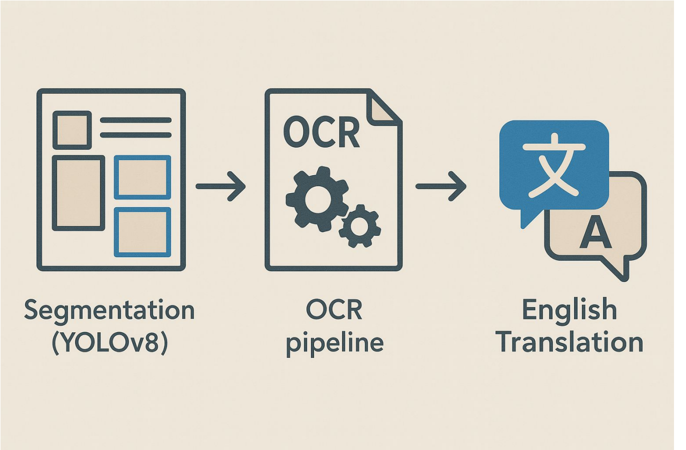
Methodology & Pipeline

Image Segmentation (YOLO-based) to distinguish ads vs. articles.

Yolov8s (which was one of the latest image segmentation model) model trained with 1024 images, has an mAP of **96.8%**.

OCR using a mix of pipelines depending on content type (articles vs ads) and language.

Translation using IndicTrans2 for Hindi/Telugu ➔ English



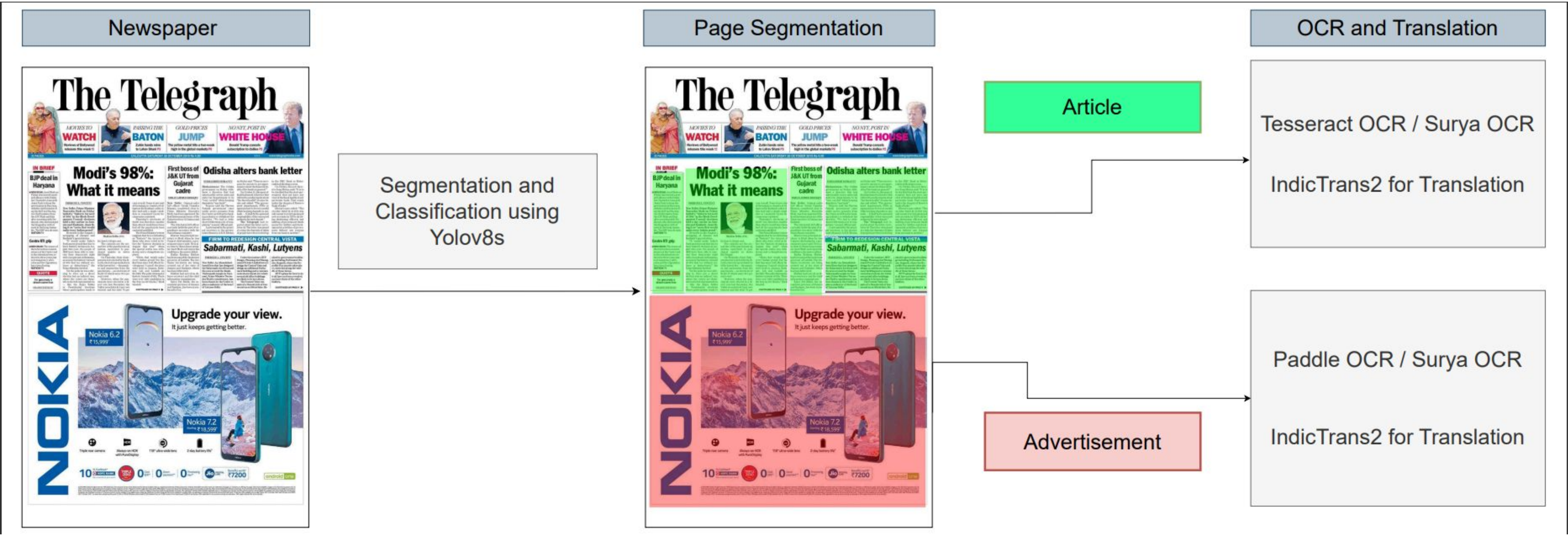
OCR Strategy

Hindi and Telugu articles were processed through SuryaOCR[1] which performed best over Indic data.

English advertisements were processed through PaddleOCR[2] which prioritised extracting maximal text in an unstructured manner.

English articles were processed through TesseractOCR[3] which also allowed extraction in reading order.

1. https://github.com/datalab-to/surya
2. https://arxiv.org/abs/2507.05595
3. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33418.pdf


Newspaper
Page Segmentation
OCR and Translation
Segmentation and Classification using Yolov8s
Article
Advertisement
Tesseract OCR / Surya OCR
IndicTrans2 for Translation
Paddle OCR / Surya OCR
IndicTrans2 for Translation

Identifying Govt. and Company Ads

Keyword based identification for Govt and Company advertisements and articles.

The Companies chosen were based on prominent advertisers[1] in newspapers. Keywords associated were based on popular brands and products from the company as well.

Keyword misclassifications are random across pages and areas, minimizing risk of systematic bias in regressions and visual analyses.

1. TAM Media Research India's quarterly reports. (https://tamindia.com/)

Identifying Govt. and Company Ads

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

Cost Calculation for Ads

Each ad's bounding box was converted to cm^2.

The cost per cm^2 of an advertisement depended on the City/ Edition, the source and the page number. Based on publicly available rate cards and the area of the advertisement and these factors, the final cost per ad were calculated.

Classified first, third, and last pages as “premium” spots; all other pages as non-premium.

Table 5.1: 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

Who Advertises?

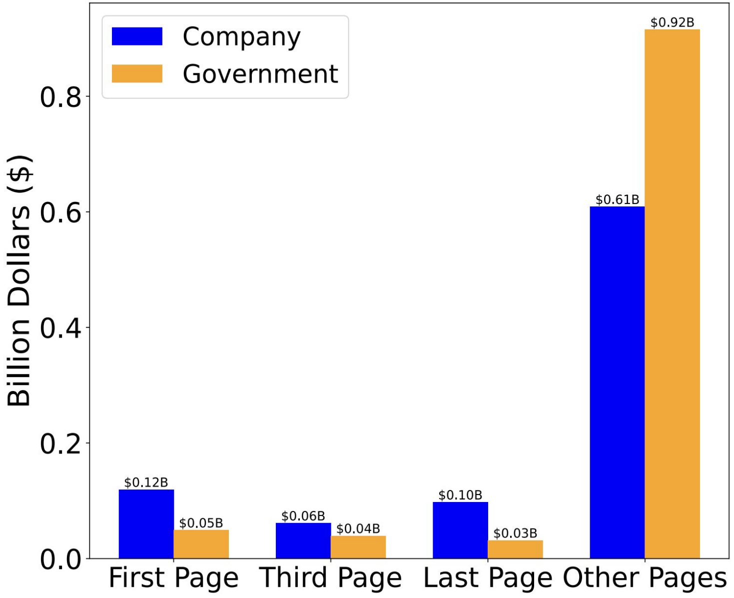
Government entities—encompassing both central and state-level agencies—allocated a total of **$1.02 billion** to mandated notices and small-format advertisements.

Private corporations—including FMCG, automobile, technology, education, and finance firms—invested a total of **$890 million** in strategic branding and large-format advertising.

Who Advertises?

Government entities—encompassing both central and state-level agencies—allocated a total of **$1.02 billion** to mandated notices and small-format advertisements.

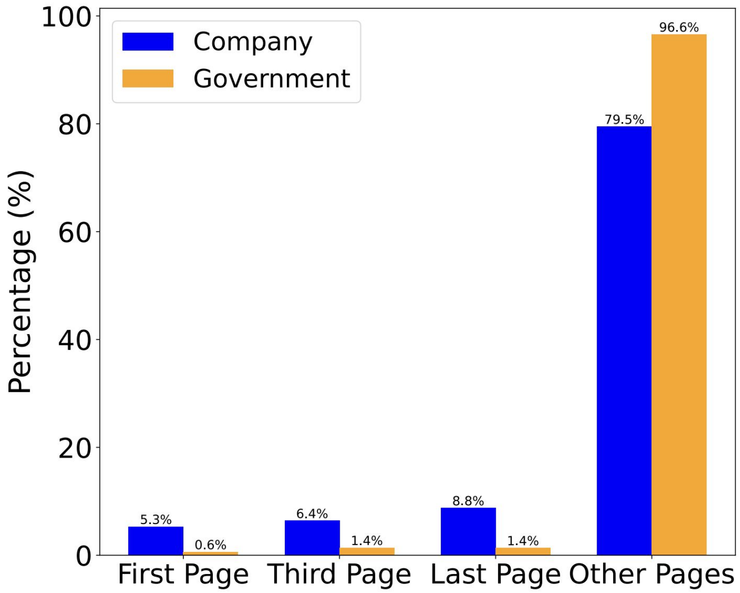
Private corporations—including FMCG, automobile, technology, education, and finance firms—invested a total of **$890 million** in strategic branding and large-format advertising.



Where Are Ads Placed?

~20% of corporate ads appear on premium pages.

~97% of government ads are placed on non-premium pages.



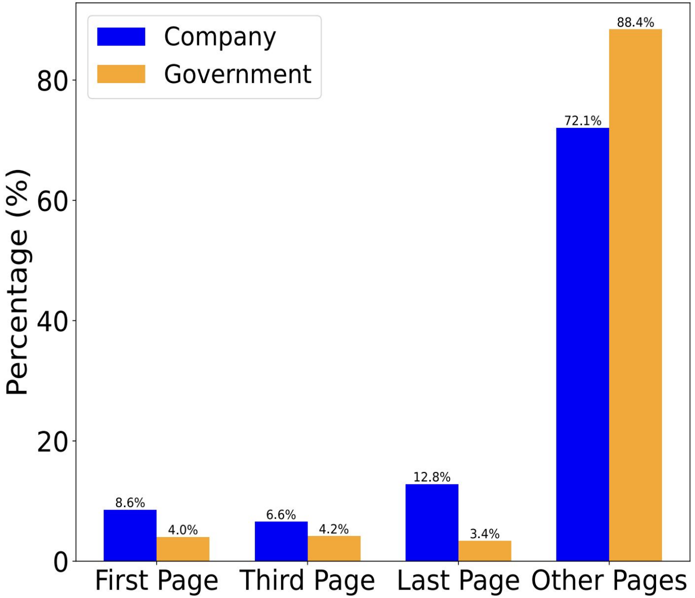
Where Are Ads Placed?

27.9% of corporate ad area appear on premium pages, capturing **31.6%** of their total print ad spend.

88.4% of government ad area is placed on non-premium pages, with only **11.6%** on premium spots.

When used, both corporate and government ads occupy over 55% of page area on premium pages.

Corporates prioritize visibility by investing nearly one-third of their budget on prime pages, whereas governments spread expenditures evenly across all pages.



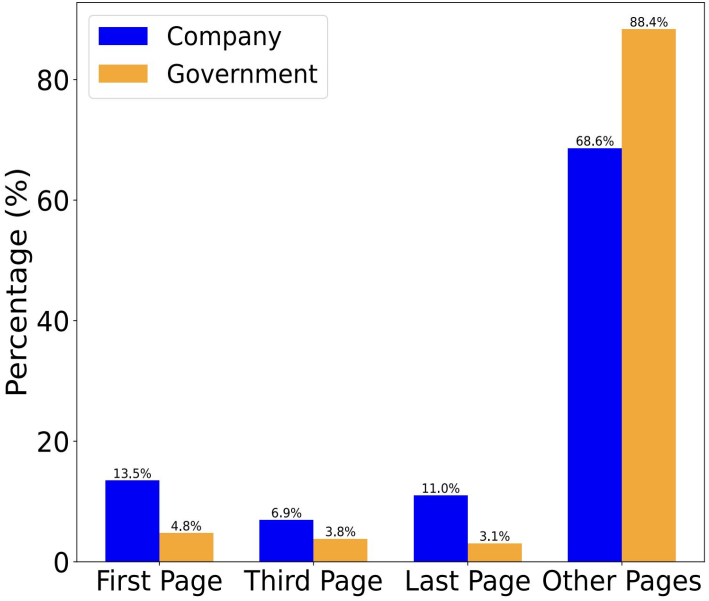
Where Are Ads Placed?

27.9% of corporate ad area appear on premium pages, capturing **31.6%** of their total print ad spend.

88.4% of government ad area is placed on non-premium pages, with only **11.6%** on premium spots.

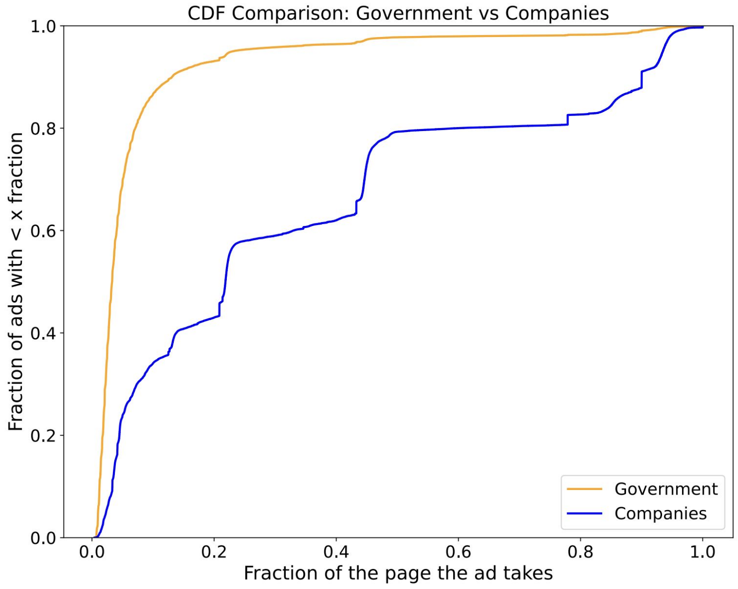
When used, both corporate and government ads occupy over 55% of page area on premium pages.

Corporates prioritize visibility by investing nearly one-third of their budget on prime pages, whereas governments spread expenditures evenly across all pages.



Design & Formatting of Ads

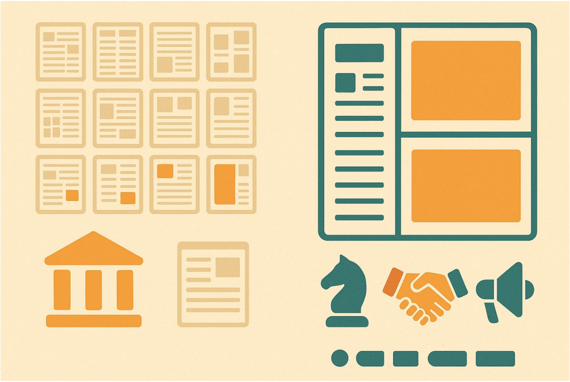
85% of government ads are small (<10% of page); corporates favor standard large blocks (25%, 50%, 100%).



Design & Formatting of Ads

85% of government ads are small (<10% of page); corporates favor standard large blocks (25%, 50%, 100%).

Companies and Industries also push a particular strategic way of advertising. Patanjali prefer Hindi newspapers than the english counterparts, Educational Institutions prefer more first page ads, and other interesting preferences.



When Are Ads Published?

Steady ~35% page area coverage annually (2018–2024).

Amid the COVID-19 pandemic, ad placements plunged sharply to around 10 % in April–May 2020, before recovering gradually by mid-2021.

Bump in ads during festive season as well from Diwali to New Year.

What is the content about?

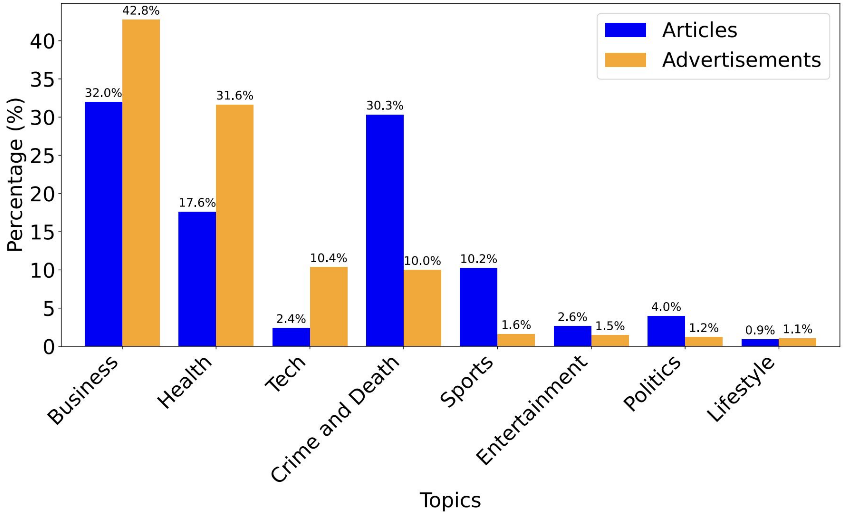
Employed an off-the-shelf RoBERTa-based transformer[1], fine-tuned on New York Times articles, to assign each ad text to one of eight high-level themes such as Business , Health, Tech, Crime, Sports, Entertainment, Politics, Lifestyle.

1. https://huggingface.co/dstefa/roberta-base_topic_classification_nyt_news

What is the content about?

Advertisements are heavily skewed toward business, health, and technology topics.

Editorial articles cover a broader range of subjects including crime, public safety, and politics.



Influence on Editorial Coverage

We construct the Weighted Ad Ratio to capture both the physical prominence and economic cost of an advertisement. It is specified as

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

And We use a panel regression with dependent variables are Sentiment and Count of Articles.

Sentiment scores are assigned using the cardiffnlp/twitter-roberta-base-sentiment-latest model. The range is from -1 for negative to 1 for positive.

Why Panel Regression

Captures variation over time and across entities (e.g., companies, newspapers), allowing for more nuanced causal inference.

Controls for unobserved, time-invariant heterogeneity (e.g., editorial leanings, company image) through fixed effects.

Ideal for studying longitudinal relationships between advertising investment and editorial coverage in the same outlet over time.

Panel Regression Framework

$$Y_{mnt} = \alpha + \beta \cdot \text{Weighted Ad Ratio}_{mnt} + \gamma_{mn} + \delta_t + \epsilon_{mnt}$$

Y_{mnt} is the dependent variable for entity n (specific company) in newspaper m at given time t.

γ_{mn} are media–entity fixed effects, controlling for time-invariant differences across newspaper–advertiser pairs.

δ_t are time fixed effects (daily/weekly/monthly), capturing common shocks.

ε_{mnt} is the idiosyncratic error term.

Results for Government Entities

In the baseline (no fixed effects), a 1% increase in the Weighted Ad Ratio is associated with a 0.073-point decline in total sentiment **(β = –0.0730, $p < 0.001$)**. After adding newspaper and time fixed effects, the effect attenuates to **β = –0.0084 ($p < 0.05$)**, but remains negative, indicating that mandated government ads do not buy more positive coverage.

Initially, government ad intensity predicts more articles **(β = 0.5810, $p < 0.001$)**, but this relationship disappears **(β = 0.0214, n.s.)** once full controls are applied, suggesting that statutory notice requirements drive placement but not additional editorial attention.

Table 4.1: 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 4.2: 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: Count of Articles				
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

Results for Corporate Advertisers

Corporates show a robust positive relationship. A 1 % rise in their Weighted Ad Ratio corresponds to a **0.0189-point** increase in sentiment ($p < 0.001$) in the simplest model, which remains significant **($\beta \approx 0.0137$, $p < 0.001$)** under newspaper×company and time fixed effects.

Higher corporate ad spend is linked to more coverage: baseline **$\beta = 0.4360$ ($p < 0.001$)**, and even with full controls **$\beta \approx 0.2232$ ($p < 0.001$)**, underscoring that strategic branding buys both volume and tone advantages

Table 4.3: 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 4.4: 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

Discussion & Implications

Commercial incentives shape editorial decisions in print media.

Corporate influence stronger than government mandates, primarily since govt revenue is ensured whereas corporate isn't.

Ethical concerns for journalistic independence and democratic discourse.

Dataset and pipeline as resources for transparency and further research.

Limitations

Reliance on OCR and keyword rules introduces some imprecision – e.g. errors in Indic-language text or dense layouts can lead to misclassified ads or partial data. These small extraction errors may propagate into the analysis.

Our regression findings are correlational – higher ad spending coincides with more positive coverage, but we cannot conclusively prove causation.

The dataset is limited to 5 sources and over a period of 5 years. Larger dataset could have more robust conclusions.

Future Work

Improving Accuracy: Using large language models to post-process OCR text could fix reading order and correct recognition errors, especially in complex ad layouts. Similarly, more advanced entity recognition could replace simple keyword matching to better identify advertisers.

Expanding Data: We consider only 5 sources over a period of 5 years, Understanding about the total landscape over a larger time period could provide more robustness and better insights regarding the results.

Better Experimental Design: With access to other features, experiments which show causality instead of correlation in our case.

More Analysis: The dataset is still underutilized, much more insights could be derived from further analysis into the dataset.

Research Contributions & Novelty

Novel Analysis Pipeline: Developed a first-of-its-kind, scalable pipeline for large-scale print news analysis (multi-language, multimodal), adaptable to other regions. This approach overcomes prior limitations in manual media studies.

Comprehensive Dataset: Assembled an unprecedented dataset of 12,000+ newspaper editions (5 major Indian dailies, 3 languages) spanning 6 years. This open dataset provides broad coverage and enables new research on media economics.

Empirical Evidence Gap: Provided strong empirical evidence that advertising spending correlates with more favorable and extensive news coverage, especially for corporate advertisers – addressing a long-debated question with data.

Cross-Disciplinary Contribution: Introduced novel computational methods to journalism studies – combining computer vision, NLP, and econometric analysis – and open-sourced the code & data for transparency and future work.

Publications

Harsha Vardhan Nemani, Ponnurangam Kumaraguru, and Kiran Garimella. 2025. “Analyzing Patterns and Influence of Advertising in Print Newspapers. In ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS ’25), July 22–25, 2025, Toronto, ON, Canada. ACM, New York, NY, USA 17 Pages. https://doi.org/10.1145/3715335.3735468

Acknowledgements

Prof PK for the support and freedom across my research journey.

Prof Kiran Garimella for mentorship along this project.

Tejasvi, Rohan and Prof Ashwin for collaborating on my first research project.

Precog and all the good folk who reviewed my content too many times to count.

Prof Radhika Krishnan and Prof Rajorshi Ray for reviewing and providing valuable feedback.