Supplementary material
MAdVerse: A Hierarchical Dataset of Multi-Lingual Ads from Diverse Sources and Categories

November 8, 2023

We provide additional insights and statistical details pertaining to our collected data, along with a comprehensive presentation of the experimental outcomes for the tasks, complemented by supplementary visualizations.

1 Examples of ads and non ads

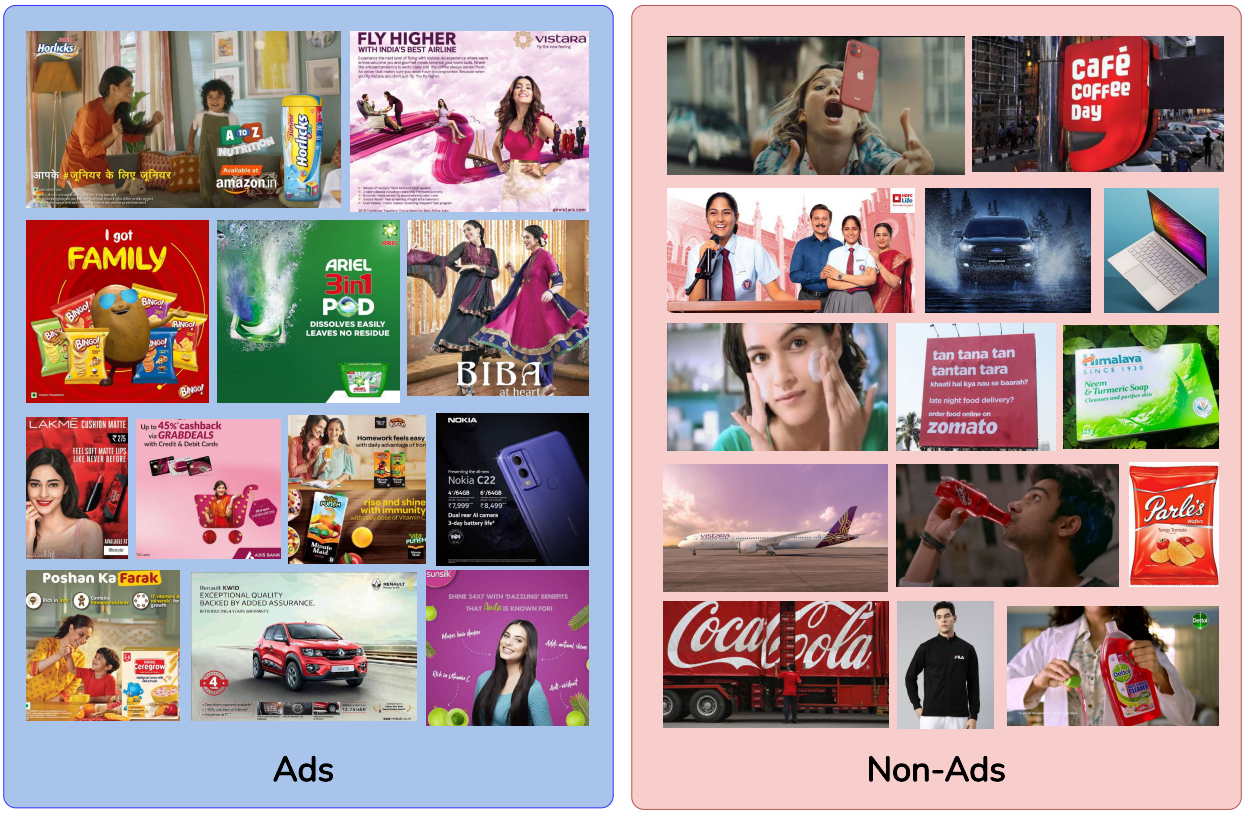

Figure 1: Sample images from ads and non ads

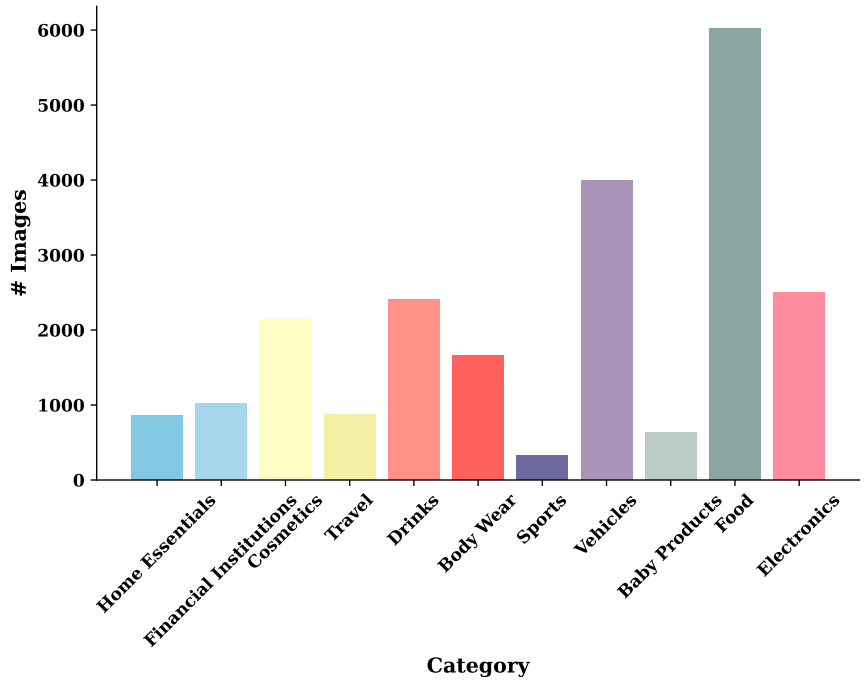

Figure 2: Number of images per category

2 Pipeline to Extract Newspaper Ads

In this section, we outline our methodology for extracting newspaper advertisements from digital newspapers. Our approach involved multiple steps, beginning with the acquisition of 30,987 ads from digital newspaper sources.

We employed a state-of-the-art Faster R-CNN model with a ResNet-50 Backbone, enhanced with a Feature Pyramid Network (FPN) architecture. This model was pre-trained on the Newspaper Navigator Dataset, setting the foundation for our subsequent fine-tuning efforts.

Fine-tuning of the network was carried out by freezing the weights of the backbone and adjusting the weights of the fully connected layers. This fine-tuning process relied on a carefully annotated dataset of 2,863 English newspaper page images, utilizing an 80-20 percent train-test split strategy.

The performance of our ad-detection model is summarized in Table 1 below:

Model	AP	AP50	AP75	Class Accuracy (%)
Faster R-CNN	71.52	78.52	74.25	92

Table 1: Performance Metrics of the Faster R-CNN Model

Subsequently, we implemented an ad classifier based on a ViT/ConvNext Backbone to further enhance the precision of our ad extraction pipeline. This classifier played a crucial role in eliminating false positives from the data obtained through the detection stage.

2.1 Dataset Statistics

We acquired a multilingual dataset comprising 29,031 images from various languages, with the following distribution:

Language	Number of Images
Gujarati	2,292
Kannada	3,145
Odia	2,073
Tamil	1,804
Telugu	2,599
Urdu	410
Hindi	9,081
Marathi	4,283
Bengali	1,891
Malayalam	1,453

Table 2: Multilingual Dataset Statistics

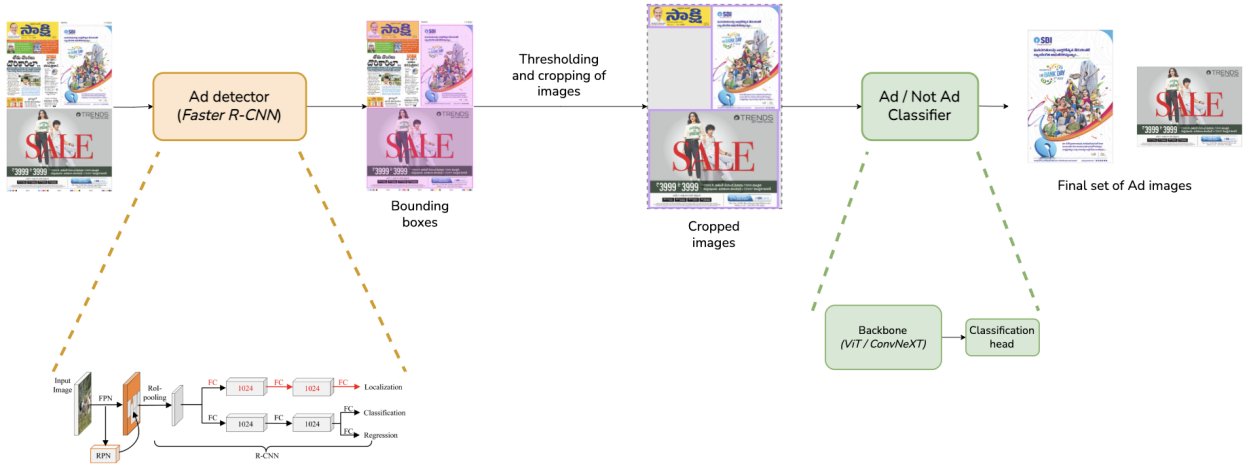

Figure 3: A Faster R-CNN-based detection pipeline designed to identify potential advertisement images, extract bounding boxes, crop images, and filter ads from non ads using an Ad/Non ad classifier

2.2 Ad Classifier Performance

We trained an ad classifier on a subset of the dataset, consisting of 22,294 advertisements and 47,192 non-advertisement samples, using three different split ratios (0.7, 0.2, and 0.1). The classifier's performance is summarized below:

Dataset	F1 Score	Accuracy (%)
validation	87.57	92
test	88.80	92.75

Table 3: Ad Classifier Performance Metrics

These statistics provide insights into the composition of our multilingual dataset and the effectiveness of our ad classifier in distinguishing advertisements from non-advertisement content.

For a visual representation of the entire pipeline, please refer to the diagram provided in Figure 3.

3 Classes present in hierarchy

The following are the names of categories, sub-categories and brand names which are organized hierarchically. We have a total of 11 categories, 51 subcategories and 524 brands. Fig 4 is a radial tree representation of the below hierarchy.

NOTE: *MISC* in each and every subcategory contains images which cannot be put under a brand in that subcategory, and it contains images which are a collaboration of multiple brands related to that subcategory.

- Baby products
  - Baby essentials
    - Himalaya baby products, MISC, Johnson & Johnson baby products, MamaEarth baby products, Moms Co. baby products, Sebamed baby products, Dove baby products, Chicco baby products
  - Baby food
    - Nestle CERELAC, MISC, Nestle CEREGROW, Slurrp baby food
  - Diapers and Wipes
    - MamyPoko diapers and wipes, Supples diapers and wipes, MISC, Huggies diapers and wipes, Himalaya diapers and wipes, Pampers diapers and wipes
- Body wear
  - Clothing brands
    - Flying Machine, Fabindia, Provogue clothing, MISC, Raymonds, Louis Philippe, Gini & Jony, Blackberry, Monte Carlo, Biba, Allen Solly, Peter England
  - Footwear
    - Nike, Puma, MISC, Skechers, Wood Land, Bata, Campus, Bahamas footwear, Reebox, Sparx, Red Chief
  - Jewellery
    - Joyalukkas jewellary, MISC, Tanishq jewellary, Malabar gold and diamond, Blue Stone jewellary, PC jewellers, Kalyan jewellers
  - Watch
    - Sonata watches, Rolex watches, Casio watches, MISC, Emporio Armani watches, Fastrack watches, Citizen watches, Titan watches, Tommy Hilfiger watches
- Cosmetics
  - Fragrances
    - Park Evenue, MISC, Denver, Fog cosmetics, Engage, Wild Stone, Axe
  - Hair care
    - Set Wet, Pantene shampoo, MISC, Head & Shoulders shampoo, MamaEarth shampoo, L'Oréal Paris shampoo, Tresemme shampoo, WOW Skin Science shampoo, Indulekha shampoo, Clinic Plus shampoo, Sunsilk shampoo

  - Sking cleansing
    * Clean and clear, MISC, Dettol soaps, Santoor soaps, Nivea soaps, Medimix soaps, Biotique soaps, Lifebuoy soaps, Dove soaps, Park Avenue soaps, Himalaya soaps, MamaEarth soaps, Fiama Di Wills soaps, Pears soaps, Cinthol soaps
  - Skincare and Makeup
    * Johnsons, Ponds, L'Oreal, Lux soaps, MISC, Maybelline, Amway, Mamaearth, Himalaya Herbals, Lotus Herbal, Olay, Lakme, Garnier, Biotique cosmetics, Fair and Lovely

- Drinks
  - Alcohol
    * Carlsberg, MISC, Kingfisher liquor and alcohol, Heineken Beer, McDowell's Liquor & Alcohol, Royal Challenge Liquour and Alcohol, Royal Stag, Old Monk Liquor & Alcohol, Officers Choice Liquour & Alcohol, Blenders Pride
  - Coffee
    * "Bru coffee", Bru Ads, Café Coffee Day, MISC, Black Baza Coffee, Flying Squirrel Coffee, Country Bean coffee, Blue Tokai coffee, Seven Beans Coffee
  - Energy drinks
    * Rockstar Energy Drink, Sting, Gatorade, MISC, Red Bull, Ocean Energy Drink, Enerzal, Monster Energy Drink, Glucon-D
  - Health beverages
    * Horlicks nutrition drink, PediaSure, Patanjali Powervita, MISC, ProtienX nutrition drink, Ensure nutrition drink, Yakult probiotic nutrition drink, Cadbury Bournvita nutrition drink, Boost nutrition drink, Complan
  - Juice
    * Minute Maid, Roohafza, MISC, Tropicana, Maaza, Drunken Monkey, Paper boat, Real Fruit Juice, Rasna, Slice, Frooti
  - Soda
    * Coca Cola, Mirinda, MISC, Fanta, Limca, Pepsi, 7UP, Mountain Dew, ThumbsUp cold drinks, Sprite
  - Packaged water
    * Kinley, Himalayan Mineral Water, MISC, Patanjali Divya jal, Smartwater, Bisleri

- Electronics
  - Air conditioners
    * Hitachi air conditioners, MISC, Blue Star air conditioners, Daikin air conditioners., Voltas air conditioners, "Samsung" air conditioners, Indina Ads, LG air conditioner
  - Computers

        * MIcrosoft laptops and computers, Asus laptops and computers, MISC, Intel laptops and computers, Dell laptops and computers, MSI laptops and computers, Lenovo laptops and computers, Sony laptops and computers laptops, Razer laptops and computers Laptops, Acer laptops and computers, Apple laptops and computers, HP laptops and computers, MI laptops and computers

    - Mobile devices
        * One-Plus mobiles, MISC, Gionee mobiles, Samsung mobiles, Motorola mobiles, Apple mobiles, LG mobiles, Lava Mobiles mobiles, Nokia mobiles, Sony mobiles, HTC mobiles, Asus mobiles, Oppo mobiles, Google Mobiles, Videocon mobiles, Vivo mobiles, Xiaomi mobiles, Micromax mobiles

    - Refrigerators
        * Panasonic refrigerators, Godrej refrigerators, LG refrigerators, Lloyd refrigerators, Whirlpool refrigerators, Haier refrigerators, Samsung refrigerators

    - Tel-com service providers
        * Airtel, BSNL, Tata Teleservices, Docomo, Reliance, MTNL, Idea, Vodafone, Jio

    - Television brands
        * Panasonic television, Philips television, Samsung television, Sony television, LG television, MI television, Sansui television, Micromax television

    - Washing machine
        * Croma washing machine, Whirlpool washing machine, MISC, Panasonic washing machine, Bosch washing machine, LG washing machine, Godrej washing machine, Haier washing machine

- Financial institutions
    - Banks
        * Union Bank Of India, Canara Bank, MISC, Bank of Baroda, Axis Bank, Bank of India, HDFC Bank, Yes Bank, State Bank Of India, Allahabad Bank, Punjab National Bank, Central Bank of India, ICICI Bank

    - Insurance
        * MISC, Bajaj Allianz, Tata AIA, Birla Sun Life, PNB Metlife, HDFC Life, Exide Life, LIC, Max Life, Kotak Life, SBI Life, ICICI Prudential

- Food
    - Baked snacks
        * Good Day biscuits, Jim Jam, PriyaGold Cheese cracker, Britannia Marie Gold, MISC, Hide & Seek, Anmol Biscuits, Dark Fantasy, Priya Gold Biscuits, 20-20 Cookies, Krackjack, Parle G, Bisk farm, Bourbon biscuits, Butter Bite biscuits, Unibic, Oreo, Monaco biscuits

    - Food brands
        * Amul, Patanjali, Britannia, MISC, Hatsun Agro, MTR Foods, KRBL Limited, Kissan, Sunfeast, Nestle, Parle Agro, Hindustan Unilever, McCain, Heritage Foods Limited

- Chewing gum
  - Boomer chewing gum, Trident chewing gum, MISC, Centre Fruit chewing gum, Orbit chewing gum, Wrigley double mint, Nicotex, Happydent chewing, Big Babol chewing gum, Centre Fresh chewing gum, Mentos
- Condiments
  - MISC, Maggi Masala, Ashok masale, Heinz, Suhana masale, Badshah masale, "Everest" masale, Ads, Goldiee masale, Chings masale, Catch masale, MDH
- Confectionery
  - Cadbury Silk chocolate, Eclairs chocolate, MISC, Cadbury Gems chocolate, Milky Bar chocolate, Barone chocolate, Kopiko chocolate, M&M's chocolate, Kisses chocolate, Munch chocolate, Cadbury 5-Star chocolate, ChocOn chocolate, Alpenlibe chocolate, Ferrero Rocher chocolate, Just jelly chocolate, Snickers chocolate, Kit Kat chocolate, Perk chocolate, Melody chocolate, Dairy Milk chocolate, Raffaello chocolate, Hershey chocolate, Kacha Mango Bite chocolate, Mars chocolate, Twix chocolate
- Delivery apps
  - Zomato, Grubhub, MISC, FreshMenu, Dunzo, EatSure, Box, Swiggy, Eatfit, Travelkhana, Deliveroo
- Dry fruits and Nuts
  - MISC, Nutraj dry fruits, Rostaa dry fruits, Happilo, Nutty Gritties, Tulsi Dry Fruits
- Eateries
  - The Thickshake Factory, Faasos Food Company, MISC, OvenStory, Bikanervala, Smokin' Joes, Sagar Ratna, Haldiram, Wow! Momo, Jumbo King, Nirula's Restaurant, Parsa's-Food For all, Shiv Sagar, Barbeque Nation, Paradise Restaurant, Saravana Bhavan, Kaati Zone, Ratna Cafe, Karims Restaurant, Chai Point
- Flour
  - Fortune Chakki Fresh Atta, Organic Tattva Whole Wheat Flour, Aashirvaad Atta, Pillsbury Chakki Fresh Atta, Ahaar Whole Wheat Atta, MISC, Nature Fresh Sampoorna Chakki Atta, Shakti Bhog Chakki Fresh Atta, Patanjali Whole Wheat Atta, Laxmi Bhog Whole Wheat Atta, Rajdhani Chakki Fresh Atta
- Fried snacks
  - Kurkure wafers, Cheetos wafers, Uncle Chips, Bingo wafers, MISC, Parle's Wafers wafers, Lay's wafers, Too Yum chips, Yellow Diamond wafers, Haldiram wafers, Crax wafers, Balaji Wafers
- Ice cream
  - Amul ice-cream, MISC, Havmor ice-cream, Kwality Wall's, Cream Stone, Mother Dairy, Naturals ice-cream, Cream Bell, Vadilal, Cream And Fudge
- Noodles
  - Knorr Soupy Noodles, Yippee Noodles, MISC, Chings Secret Noodles, "Patanjali" Noodles, indian Ads, Top Ramen Noodles, Wai Wai Noodles

- Home essentials
  - Dental care
    - Pepsodent toothpaste, Dabur Red toothpaste, MISC, Close-up toothpaste, Colgate toothpaste, Oral-B toothpaste, Meswak toothpaste, Patanjali Dant Kanti toothpaste, Vicco Vajradanti ayurvedic toothpaste, Sensodyne toothpaste
  - Detergents
    - Surf Excel detergent, Henko detergent, MISC, Rin detergent, Tide detergent, Ariel detergent, Ghadi detergent
  - Disinfectants
    - Dettol floor cleaner, Savlon disinfectant
  - Toilet and Floor cleaners
    - Lizol disinfectant, MISC, Harpic toilet cleaner, Sani Fresh toilet cleaner, Domex floor and toilet cleaner
- Sports
  - Sports apparel
    - Sareen Sports, MISC, HRX, Adidas, Nivia, Tyka, FILA
  - Sports equipment
    - MRF, Cosco Sports
- Travel
  - Airlines
    - Vistara, Air India, Jet Airways, SpiceJet, Indigo, GoAir
  - Resorts
    - Niraamaya Retreats Backwaters And Beyond, MISC, Radisson Hotel, ITC hotels, The Oberoi Groups, Club Mahindra, The Naini Retreat, "The Roseate", Indain Ads, Le Meridien Group of Hotels, Jaisalmer Marriott Resort & Spa
  - Travel connect
    - Make MyTrip, MISC, ClearTrip, "The Travel Guru", Travel Ads, Yatra, EaseMyTrip, OYO, "RedBus", Bus Ads, Goibibo
- Vehicles
  - Four Wheeler
    - Bajaj
      - Bajaj Qute
    - Car brands
      - Jaguar cars, Maruti Suzuki, Mahindra cars, Fiat cars, MISC, Rolls Royce, Ford cars, Tata Motors, Skoda cars, Nisson cars, Renault cars, Honda Motor Company, BMW, Volkswagen, Hyundai, Toyota

    - Ford
        - MISC, Ford Fiesta, Ford Ecosport, Ford Aspire, Ford Endeavour, Ford Figo
    - Honda
        - MISC, Honda Jazz, Honda Amaze, Honda City, Honda WR-V
    - Hyundai
        - Hyundai Santro, Hyundai i20, Hyundai Creta, MISC, Hyundai Venue car, Hyundai Alcazar, Hyundai i10, Hyundai Verna, Hyundai Tucson
    - Mahindra
        - MISC, Mahindra Thar, Mahindra Alturas, Mahindra XUV, Mahindra Marazzo, Mahindra KUV, Mahindra Scorpio, Mahindra Bolero
    - Maruti Suzuki
        - Maruti Suzuki Ignis, Maruti Suzuki Wagon R, MISC, Maruti Suzuki Alto, Maruti Suzuki Celerio, Maruti Suzuki Baleno, Maruti Suzuki Dzire, Maruti Suzuki Brezza, Maruti Suzuki Swift, Maruti Suzuki Ertiga, Maruti Suzuki Ciaz
    - Renault
        - Renault Triber, MISC, Renault Kiger, Renault KWID
    - Skoda
        - Skoda Kushaq, MISC, Skoda Octavia, Skoda Kodiaq, Skoda Slavia
    - Tata
        - Tata Nexon, MISC, Tata Tiago, Tata Tigor, Tata Safari, Tata Harrier, Tata Altroz, Tata Punch
    - Toyota
        - MISC, Toyota Fortuner, Toyotal Land Cruiser, Toyota Corolla, Toyota Innova
- Two wheeler
    - Bajaj
        - Bajaj Dominar, Bajaj Pulsar, MISC, Bajaj Platina, Bajaj Avenger
    - Motorcycle brands
        - Bajaj, TVS Motor Company, Hero MotoCorp
    - Hero Motorcorp
        - Hero Splendor, Hero Maestro, Hero Passion Pro, MISC, Hero Pleasure, Hero HF Deluxe, Hero Glamour
    - TVS
        - TVS Jupiter, MISC, TVS Raider, TVS Scooty Zest, TVS Sport, TVS Apache, TVS Scooty Pep
- Tyres
    - MRF Tyres

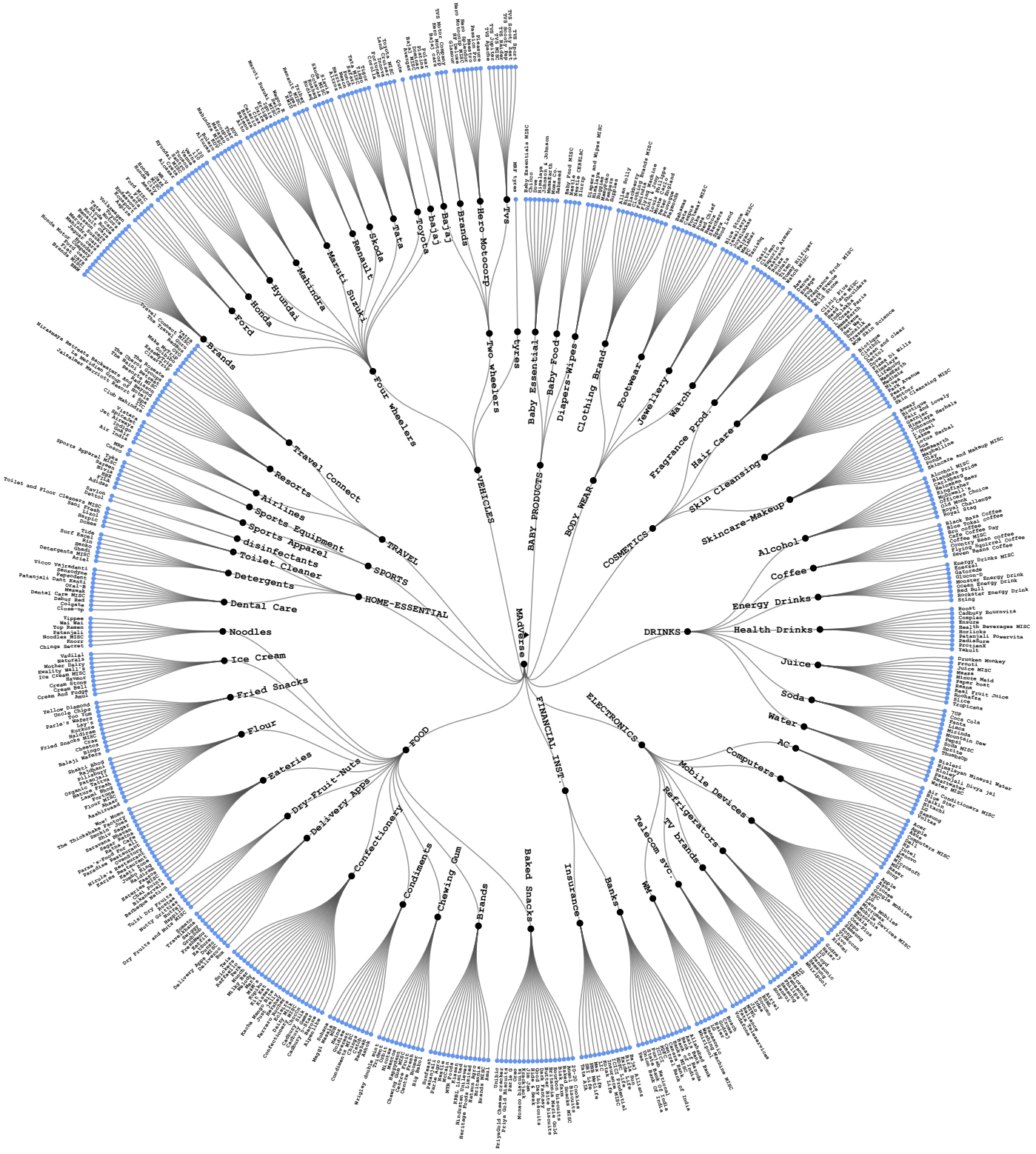

Figure 4: A radial tree visualization depicting the intricate hierarchical arrangement present in the dataset.

4 Hierarchical classification

backbone	loss	feature fusion	accuracy (%)			Height of LCA			Height of LCA (mistakes only)			TIE			TIE (mistakes only)		
			L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
BLIP-2	Sum CE	no	95.51	89.34	80.87	0.04	0.15	0.31	1.00	1.41	1.61	0.09	0.30	0.62	2.00	2.83	3.23
ConvNeXT-L			88.41	79.32	57.25	0.12	0.31	0.74	1.00	1.52	1.74	0.23	0.63	1.49	2.00	3.04	3.48
ViT-L			78.33	63.69	37.35	0.22	0.57	1.21	1.00	1.57	1.93	0.43	1.14	2.42	2.00	3.13	3.86
BLIP-2		yes	95.16	89.25	81.73	0.05	0.15	0.29	1.00	1.41	1.60	0.10	0.30	0.58	2.00	2.82	3.20
ConvNeXT-L			87.99	79.26	60.62	0.12	0.31	0.68	1.00	1.48	1.72	0.24	0.62	1.36	2.00	2.97	3.45
ViT-L			76.56	62.96	40.03	0.23	0.57	1.14	1.00	1.55	1.91	0.47	1.15	2.29	2.00	3.10	3.81
BLIP-2	DOT	yes	89.47	29.68	2.23	0.11	0.81	1.76	1.00	1.15	1.80	0.21	1.62	3.53	2.00	2.30	3.61
ConvNeXT-L			83.79	12.58	2.79	0.16	1.03	1.88	1.00	1.18	1.94	0.32	2.07	3.77	2.00	2.37	3.87
ViT-L			75.83	25.79	2.03	0.24	0.98	1.91	1.00	1.32	1.95	0.48	1.97	3.82	2.00	2.65	3.90
BLIP-2		no	81.16	28.68	1.55	0.19	0.90	1.99	1.00	1.26	2.02	0.38	1.80	3.98	2.00	2.53	4.05
ConvNeXT-L			83.57	30.25	5.42	0.16	0.89	1.80	1.00	1.28	1.90	0.33	1.79	3.60	2.00	2.56	3.80
ViT-L			75.37	25.83	1.77	0.25	0.98	1.92	1.00	1.33	1.95	0.49	1.97	3.84	2.00	2.65	3.91

Table 4: Results of all configurations within the multi-level hierarchical classifiers, encompassing both feature fusion and non-feature fusion scenarios, along with diverse combinations of backbones and loss functions.

backbone	loss	accuracy (%)			Height of LCA			Height of LCA (mistakes only)			TIE			TIE (mistakes only)		
		L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
BLIP-2	Cross Entropy	96.22	91.09	80.43	0.04	0.13	0.32	1.00	1.42	1.65	0.08	0.25	0.65	2.00	2.85	3.30
	HXE	96.29	91.55	80.58	0.04	0.12	0.32	1.00	1.44	1.63	0.07	0.24	0.63	2.00	2.88	3.25
	DOT	68.09	25.85	1.77	0.32	1.06	2.04	1.00	1.43	2.08	0.64	2.12	4.09	2.00	2.86	4.16
	Soft labels	97.46	93.74	85.25	0.03	0.09	0.24	1.00	1.41	1.60	0.05	0.18	0.47	2.00	2.81	3.19
	Semantic embeddings	98.17	94.56	52.54	0.02	0.07	0.55	1.00	1.34	1.15	0.04	0.15	1.09	2.00	2.67	2.31
ConvNeXT-L	Cross Entropy	89.41	80.27	60.15	0.11	0.30	0.70	1.00	1.54	1.76	0.21	0.61	1.40	2.00	3.07	3.52
	HXE	89.83	80.72	59.95	0.10	0.29	0.70	1.00	1.53	1.74	0.20	0.59	1.39	2.00	3.06	3.47
	DOT	83.48	34.92	2.81	0.17	0.82	1.79	1.00	1.25	1.84	0.33	1.63	3.58	2.00	2.51	3.68
	Soft labels	90.45	82.57	64.49	0.10	0.27	0.62	1.00	1.55	1.76	0.19	0.54	1.25	2.00	3.10	3.52
	Semantic embeddings	91.15	78.73	20.23	0.09	0.30	1.10	1.00	1.42	1.38	0.18	0.60	2.20	2.00	2.83	2.76
ViT-L	Cross Entropy	79.99	66.23	42.13	0.20	0.54	1.12	1.00	1.59	1.93	0.40	1.08	2.23	2.00	3.19	3.86
	HXE	80.94	66.25	41.13	0.19	0.53	1.12	1.00	1.56	1.90	0.38	1.06	2.23	2.00	3.13	3.79
	DOT	75.41	29.28	3.03	0.25	0.95	1.92	1.00	1.35	1.98	0.49	1.91	3.85	2.00	2.70	3.97
	Soft labels	80.83	67.23	43.03	0.19	0.52	1.09	1.00	1.59	1.91	0.38	1.04	2.18	2.00	3.17	3.82
	Semantic embeddings	82.42	64.33	10.2	0.18	0.53	1.43	1.00	1.49	1.59	0.35	1.07	2.86	2.00	2.99	3.19

Table 5: Results of all the variations of leaf-only hierarchical classifiers, employing diverse combinations of backbones and loss functions.

BLIP-2 stands out for its accuracy across all levels, regardless of the type of loss or feature fusion. Check out Figure 5 for trends. The DOT loss holds its own at coarser levels but struggles as things get more detailed. It's not a reliable choice for fine-grained classification in our dataset hierarchy

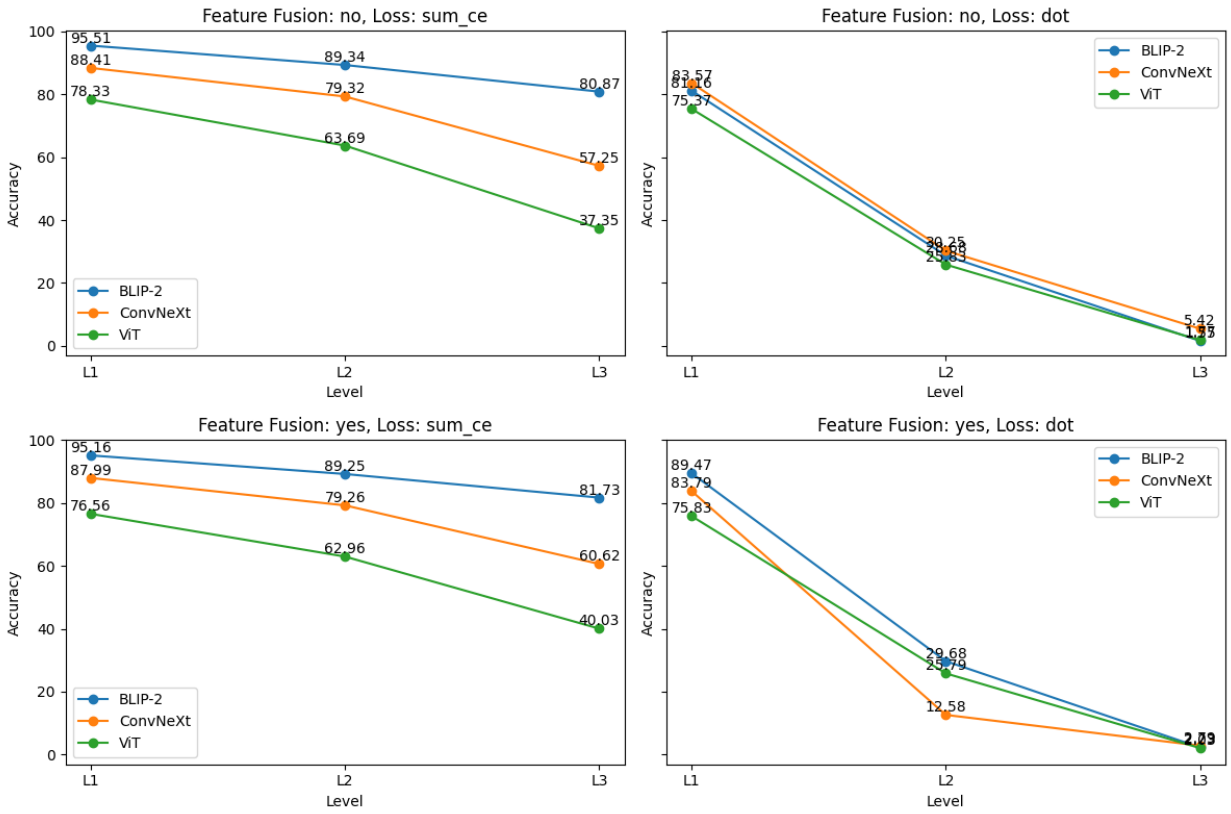

Figure 5: Trends in level wise accuracy in multilevel architectures with varied loss functions and feature fusion status

5 Metrics and losses

5.1 Metrics

Two classes of metrics were used. One is hierarchy agnostic, and the other is hierarchy aware. **Hierarchy agnostic**: These are standard metrics like level-wise accuracy. They don't take into account the hierarchical structure of the data and treat all levels equally..

Hierarchy aware: We have used 4 types of hierarchy aware metrics:

- Height of LCA: LCA represents the common ancestor at the lowest level connecting predicted and ground truth nodes. Higher values of this metric indicate errors in the upper levels of the hierarchy.

- Height of LCA (mistakes only): Similar to the above metric but focused specifically on misclassifications. This metric helps us assess the severity of mistakes made by the hierarchical classifier.

- TIE (Tree Induced Loss):TIE measures the shortest distance between the predicted and ground truth nodes. When both nodes are on the same level, it simplifies to twice the LCA height. In imbalanced hierarchies, it represents a distance from the ground truth to predicted node, indicating how far the predictions have deviated.

- TIE (mistakes only):This metric is the same as TIE but only considers misclassifications. It provides

insight into the severity of mistakes made by the hierarchical classifier.

For all hierarchy-aware metrics, lower values indicate better performance

5.2 Losses

This is a brief overview of the loss function we used.

- Sum CE: Sum of cross entropy loss from all levels, which is like treating all the levels individually and equally without taking the hierarchy into account.
- Simple CE: Just a simple cross entropy loss
- HXE: This is hierarchical cross entropy, which takes level wise softmax probabilities and uses conditional probability to encode the information about the hierarchy.
- Soft labels: is a label-embedding approach, which uses a mapping function $y()$ to associate classes with representations which encode class-relationship information that is absent in the trivial case of the one-hot representation.
- Semantic Embedding: This maps all the one-hot targets onto a unit-hyper-sphere of the same dimension as the number of classes, enforcing tree based distances between the vectors to make the target embedding hierarchy aware.
- DOT loss: It uses tree based metrics to construct a matrix called ground distance matrix, which is used in discrete optimal transport framework.

5.3 Source classification

In our dataset, the images are accompanied by annotations indicating their sources. We can broadly categorize the data into two main groups based on their origin. The first group comprises advertisements that were extracted from traditional newspapers. Additionally, this group includes ads obtained from various online platforms, which we refer to as "web ads."

The second grouping encompasses advertisements obtained from multiple sources. It includes newspaper ads as well as advertisements found on social media platforms such as Facebook and Instagram. Lastly, this group comprises ads that were scraped from the Google database, commonly known as "Google ads." These two distinct groupings help us analyze the diverse sources of advertising content within our dataset.

The TSNE plots are shown in the figure 6 for both types of groupings. Based on the plotted data, it's evident that Google ads and social media ads exhibit a high degree of similarity, as their data points cluster closely together. This suggests that the images from these sources share common characteristics.

In contrast, the data points for newspaper ads are more dispersed across both sources, indicating a mixed distribution. However, a significant portion of web ad data points appears to concentrate on the outer regions of the plot. This observation suggests that web ads and newspaper ads may be distinguishable from each other when considered in higher-dimensional spaces.

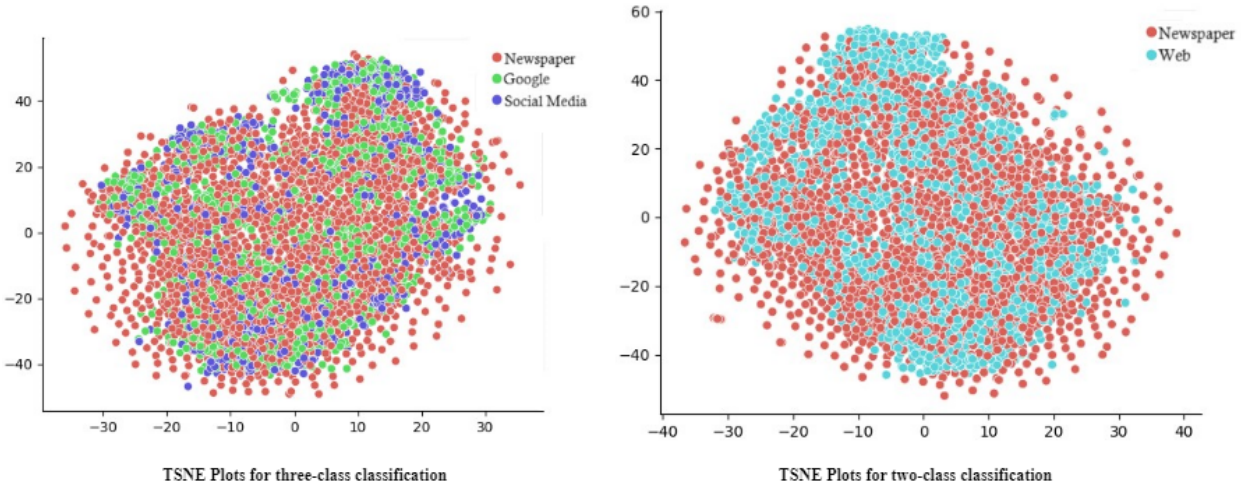

Figure 6: TSNE plots for different groupings of the source of ads.