Advertisers, Provenance, and Policy: A 30-Country Audit of Children's YouTube Ads

Anonymous Author(s)

Abstract

We present the first systematic advertiser-level audit of children's video advertising, analyzing 22,760 ad impressions from 2,928 advertisers across 30 countries. Advertiser characteristics, verification status, geographic origin, and cross-border targeting emerge as critical yet overlooked predictors of ad safety: in regions with weaker online child-protection policy, unverified advertisers are 3–4× more likely to serve age-inappropriate ads, and specific source countries (notably Ecuador and Lebanon) disproportionately supply age-inappropriate ads to vulnerable Arab and South Asian audiences, pointing to advertiser-focused interventions as essential. We also provide cross-regional evidence linking policy to safety outcomes: children in regions with weaker regulation face ∼3.6× more inappropriate exposure, reflecting weaker verification enforcement, language-specific risks, and regulatory gaps. Finally, we demonstrate that lightweight machine learning models trained exclusively on readily available ad metadata, including categories, language, location data, and advertiser attributes, attain 83.8% accuracy for automated detection of child-inappropriate ads, suggesting that metadata-based approaches can complement or reduce reliance on computationally expensive content analysis. To support reproducible research, we release a corpus of 5,611 video ads with qualitative labels for age appropriateness and relevance.

1 Introduction

YouTube has become one of the primary platforms for children's media consumption, with billions of hours of content viewed annually [51]. As children increasingly consume video online rather than through traditional broadcast media, they encounter an advertising ecosystem that is algorithmically curated, globally distributed, and inconsistently regulated [5, 10, 41]. Unlike traditional children's television with established advertising standards and regulatory frameworks, these digital platforms present unique challenges for ensuring age-appropriate commercial content [7]. The challenge is particularly acute because children's developing cognition limits their ability to recognize persuasive intent or critically evaluate advertising messages, making them especially vulnerable [32, 43]. Moreover, inappropriate ads can normalize unsuitable behaviors, expose children to adult themes prematurely, and direct them toward harmful products or services [27, 52].

These risks affect millions of children worldwide, as YouTube serves a global audience spanning diverse cultural, linguistic, and regulatory contexts [11, 51]. While the platform has implemented safety measures, including restricted advertising categories and algorithm-based classification, their effectiveness varies significantly across regions [19, 38, 47]. The platform's global reach enables advertisers from any jurisdiction to potentially reach children anywhere, creating regulatory arbitrage opportunities for less scrupulous actors to exploit enforcement gaps [45, 55]. The sheer content volume makes comprehensive manual review impractical, requiring automated systems whose performance and consistency across contexts remain questionable [28, 36]. As regulatory frameworks intensify, COPPA in the United States, GDPR in Europe, and emerging child safety legislation elsewhere, understanding the nature, distribution, and determinants of inappropriate advertising becomes crucial for platforms, policymakers, and child advocacy organizations [13, 14, 41].

Prior research has examined content moderation challenges and online risks for children but significant gaps remain in understanding systematic inappropriate advertising patterns across policy environments [2, 4, 30, 44, 48, 50]. Existing work has focused primarily on content analysis within single jurisdictions or platform-wide policies [35, 53], overlooking how advertiser verification and geographic policy variations create differential exposure risks. Non-native advertisers, particularly those targeting countries from abroad, and their relationship to ad safety remain underexplored. Furthermore, limited attention has been paid to how advertiser verification status correlates with content appropriateness. The interaction between language, content type, geographic targeting, and policy enforcement in shaping children's ad exposure remains poorly understood.

This work addresses these gaps through a systematic investigation of the nature and distribution of inappropriate ads on children's content on YouTube across diverse policy regions to identify patterns, risk factors, and potential intervention points. We pursue three goals: first, quantifying disparities in inappropriate ad exposure across different policy environments; second, examining how advertiser characteristics, particularly verification status and geographic origin, influence ad appropriateness; Third, we train lightweight machine learning classifiers exclusively on ad metadata features (categories, language, location, advertiser attributes) for automated ad-appropriateness detection, evaluate their classification performance, and identify the most influential signals through feature importance analysis.

To address these questions, we conduct a large-scale empirical study of advertising on YouTube children's content across 32 countries spanning diverse policy environments. We analyze 22,766 video ad impressions (cleaned from 26,737 collected) comprising 4,840 unique ads from 2,928 distinct advertisers operating across 101 advertising locations. These ads were displayed on 2,172 popular child-oriented videos (collectively accounting for over 445 billion views) watched in each country. Our sample includes countries from all four quartiles of the Child Online Safety Index (COSI) [29], ranging from high-regulation environments like the United States, Germany, and Singapore to lower-policy regions such as Bangladesh, Ghana, and Algeria. We manually annotate a stratified sample of 5,611 ads (up to 200 per country) using YouTube's stated child-safety policies to classify content as inappropriate, child-directed, or irrelevant for children, achieving substantial inter-annotator agreement (Cohen's $\kappa = 0.79$). For each ad, we also collect advertiser metadata including verification status, geographic location, and whether

advertisers target audiences outside their home country, enabling systematic analysis of how advertiser characteristics correlate with ad appropriateness across different regulatory contexts.

Our analysis reveals four key findings for platform governance and child protection. First, inappropriate ad exposure increases substantially from high-policy to low-policy regions, with lower-quartile countries exhibiting 3.6× more inappropriate ads and strong correlation with COSI scores, indicating systematically higher risks for children in weaker regulatory environments. Second, music videos constitute the largest source of inappropriate content (50.9%), particularly in Punjabi, Hindi, and Arabic languages, with themes unsuitable for children. Third, unverified advertisers are 3-4× more likely to serve inappropriate content in low-policy regions, suggesting mandatory verification as a high-leverage intervention. Fourth, non-native advertiser prevalence is highest in high-policy regions (69.3% in Quartile A) but decreases in lower-policy regions (22.8% in Quartile D), with Ecuador and Lebanon serving as disproportionate sources of inappropriate music-video ads targeting Arab and South Asian audiences (Table 1).

Taken together, we make three key contributions in this paper:

- **First systematic advertiser-level analysis of children's video ads:** Using the largest cross-country dataset (4,840 ads from 2,928 advertisers across 32 countries), we demonstrate that advertiser characteristics, verification status, geographic origin, and cross-border targeting, are critical but overlooked predictors of ad safety. Unverified advertisers are 3-4× more likely to serve inappropriate content in low-policy regions, while specific locations (Ecuador, Lebanon) disproportionately source inappropriate music-video ads targeting vulnerable audiences. This establishes advertiser-level analysis as essential for child safety interventions.
- **Cross-regional evidence linking policy to safety outcomes:** Children in lower-policy regions face 3.6× more inappropriate ad exposure across all four COSI quartiles, with systematic disparities driven by weaker verification enforcement, language-specific content risks, and regulatory gaps. This finding extends and reinforces evidence from an earlier study of an 8-country analysis that policy strength affects inappropriate exposure [31].
- **Automated detection feasibility:** We train lightweight metadata-only classifiers (Random Forest, Logistic Regression, CatBoost) on ad metadata including advertiser verification status, ad categories, language, geographic provenance, and COSI quartile. The best model (Random Forest) achieves 83.8% accuracy and 75% F1 for detecting child-inappropriate ads, with feature importance highlighting music-related signals, Arabic/Punjabi, verification, and provenance. This demonstrates that metadata-based approaches can complement or replace computationally expensive content analysis for scalable, advertiser-aware moderation.
- **Dataset for future research:** We use a comprehensive coding scheme to manually classify the age appropriateness and relevance of video ads qualitatively, providing a total of 5,611 tagged video ads for use in future work for inappropriate ad detection.

The dataset and source code used for our analysis and classifier are publicly available at an anonymized repository [3].

2 Methodology

In this section, we first detail our selection of child-oriented videos for collecting video ads and regions to assess cross-regional variations in ad content. Next, we describe our methodology for automating ad collection and present the annotation protocol subsequently used to label ads as (i) inappropriate, (ii) child-directed, or (iii) irrelevant for children using YouTube's child-safety guidelines.

2.1 Video Dataset Construction

We analyze in-stream video ads (skippable and non-skippable) displayed on long-form videos before, during, or after the main video, as they represent the most immersive ad format offering audio-visual content and generate the majority of advertiser spend compared to static banner ads [1]. To construct our dataset of video ads displayed on child-oriented videos, we use data from Social Blade [6], a YouTube-certified analytics platform that maintains a ranked list of channels tagged "made for kids" by total lifetime views since YouTube's Data API does not provide a default way to filter for "made for kids" videos. From the top 100 channels on this list, we query the YouTube API v3 [23] to retrieve each channel's 25 most popular non-Shorts videos. We exclude Shorts, short-form vertical videos, as they do not display in-stream ads; rather, advertisements appear as standalone Shorts within the user's feed through a separate delivery mechanism. To identify Shorts, we probe the video `/shorts/video_id` URL and classify videos based on the HTTP response status; channels consisting solely of Shorts or livestreams are then excluded. This filtering reduces are initial pool from a potential 2,500 videos to 2,172 unique videos, representing more than 445 billion views across 94 channels. This sampling captures a broad spectrum of content popular with young audiences worldwide, though regional popularity may vary.

2.2 Region Selection

To create a representative sample of 32 countries for cross-regional analysis, we used the Child Online Safety Index (COSI) score, which evaluates child cyber safety across six dimensions and assigns countries to four quartiles ranging from A (highest safety) to D (lowest safety) [29]. A higher COSI score indicates a stronger framework for child cyber safety.

To create a representative sample of 32 countries for cross-regional analysis, we employed the Child Online Safety Index (COSI) score, which evaluates child cyber safety across six dimensions and assigns countries to four quartiles ranging from A (highest safety) to D (lowest safety). A higher COSI score indicates a stronger framework for child cyber safety. We selected 8 countries from each COSI quartile by prioritizing those with the highest YouTube penetration rates [46], focusing our analysis on regions with significant platform exposure. We used NordVPN to proxy requests and watch videos from the selected regions, given its coverage of 118 countries worldwide [40]. Regions where a NordVPN endpoint was unavailable or where VPN IP ranges were blocked by YouTube were excluded. Our final sample comprised the following countries by COSI quartile:

- **Quartile A (Highest Index)**: Singapore, Taiwan, South Korea, Germany, Spain, Canada, Sweden, United States

Table 1: Key insights from our measurement study.

Key Insights	Description
(1) Significant disparity in inappropriate ad exposure across policy regions	Inappropriate ad rates increase from 15.3% in high-policy regions (Quartile A) to 55.2% in low-policy regions (Quartile D), a 3.6× increase demonstrating significant policy-dependent variation (Figure 1).
(2) Advertiser verification status correlates with ad safety	Unverified advertisers show higher inappropriate ad rates (35.3%) than verified ones (29.9%). This correlation strengthens in low-policy regions (Quartiles C and D), where unverified advertisers are 3-4 times more likely to serve inappropriate content than in Quartile A, suggesting verification status is a more salient safety indicator under weaker regulatory oversight (Figure 4).
(3) Inappropriate ads are concentrated in music videos, often in certain languages	Music videos make up 50.9% of inappropriate ads, largely Indian or Arabic content in Punjabi, Hindi, and Arabic, featuring themes unsuitable for children. This concentration suggests systematic content moderation challenges in non-English media (Figure 2, Figure 3).
(4) Dominance of non-native advertising with specific countries as key sources of inappropriate ads	Non-native advertisers account for 69.3% of ads in high-policy (Quartile A) but only 28.1% in low-policy (Quartile D) regions. Ecuador and Lebanon emerge as key sources of inappropriate music-video ads targeting Arab and South Asian audiences (Figure 5, Figure 6).
(5) Metadata-only automation offers scalable detection of inappropriate ads	Lightweight classifiers using only ad metadata (e.g., verification status, language, provenance) achieve 84.9% accuracy, showing that metadata-based approaches can complement computationally expensive content analysis (Figure 7).

- **Quartile B (Second Highest index)**: United Arab Emirates, France, Belgium, Vietnam, Ireland, Mexico, Turkey, Pakistan
- **Quartile C (Third Highest Index)**: Bahrain, Lebanon, Lithuania, Latvia, Netherlands, Austria, Estonia, Morocco
- **Quartile D (Lowest Index)**: Bulgaria, Algeria, Jordan, Sri Lanka, Bangladesh, Senegal, Ghana, Mozambique

The same video dataset is watched across regions to collect ads for comparability.

2.3 Ad Collection

Next, we scrape video and watch-feed ads displayed while watching the "made for kids" videos across the selected regions.

Automated Collection Process. To automate this process, each collection thread launches a new, logged-out Chrome window to avoid the influence of prior history, cookies, or user data on the ads served. During playback, we detect when an advertisement is shown and record its unique YouTube-assigned video ID, along with any external links or end-screen recommendations displayed. Since some videos were excessively long (e.g., over 10 hours), we limit viewing to a maximum of 20 mins per video to ensure collection scalability; 75.4% of videos in our dataset were under 20 mins. We also navigate to the "My Ad Center" option for each ad, which opens a pop-up displaying advertiser details from YouTube's Ad Transparency Center [17]. From this pop-up, we extract (i) the advertiser's name, (ii) the advertiser's location, and (iii) the advertiser's Google verification status. Finally, we query the YouTube Data API to retrieve video ad metadata, including category, language labels, and duration.

Data Standardization. The experiment was repeated for every video that failed to load across regions, after which the dataset was standardized for analysis by retaining only the common set of videos successfully processed across all regions. Of the 2,172 videos in our corpus, 101 were excluded from the analysis for all countries, as these videos either required sign-in or were regionally unavailable. After processing 719 videos each from Mozambique and Canada, we observed zero ads in the former and only three in the latter, all on a single video. We subsequently confirmed that YouTube has not enabled monetization in Mozambique [54]. For Canada, we repeated the experiment on the same 719 videos and detected the identical three ads on the same video, confirming reproducibility. We therefore excluded both countries: Mozambique as a non-monetized market, and Canada as a statistical outlier with insufficient ad observations for meaningful analysis.

Data Cleaning and Preprocessing Using this methodology, we collected 26,737 video ad impressions and applied two filtering criteria. First, we removed 2,153 impressions lacking advertiser metadata, and second, 1,847 impressions with missing Ad IDs—both due to failed extraction from JavaScript-rendered DOM elements and dynamically-loaded content. The final dataset comprised 22,766 video ad impressions representing 4,840 unique video ads from 30 countries. The final dataset retained 85.1% of collected impressions with complete metadata. Unique ads appeared multiple times across collection sessions (median: 2 impressions per ad), providing redundancy for frequently served content. Table 5 provides a detailed breakdown of this collection by country. We find substantial disparities in the frequency of all ad formats across regions, which may be explained by variations in YouTube's digital ad spending, as previously documented in the literature [31, 49].

2.4 Annotation and deductive coding

We employed deductive coding with predefined tags to classify advertisement age-appropriateness for children under 13 [26]. To minimize subjective bias, we derived our codebook directly from YouTube's globally-enforced child safety policies [19]. The complete codebook is available in the Appendix (see Tables 7 and 6). Using this coding scheme, each video received one of the following primary tags, following a hierarchical procedure: we first check whether the video is inappropriate for children, then whether it is child-directed, and finally label it as irrelevant if neither applies.

- **Inappropriate for children:** Ads containing restricted or prohibited content according to YouTube's child-safety guidelines.
- **Child-directed:** Content containing features or appeals specifically targeted at children under 13 (e.g., child-oriented toys, cartoon characters, or child-rated movies).
- **Irrelevant:** Ads that promote products or services intended primarily for audiences aged 13 and above.

For each assigned primary tag, we also assign a secondary tag that documents the classification rationale behind the choice of primary tag. Secondary tags for inappropriate content were derived from YouTube's guidelines on restricted content for children, while those for child-directed and irrelevant ads were informed by prior

work [19, 31]. For comparability across regions, we draw a random sample of 200 ad videos from each of the 30 regions, with the selection weighted by ad appearance frequency across videos. For 5 countries in our dataset, the total number of ad impressions was less than 200; thus, the maximum of what was available was collected from these countries (Table 5), leading to an eventual dataset of 5,611 ad videos.

Two annotators (ages 21-22, both fluent in English, Hindi, Urdu, and Punjabi) independently coded each advertisement by retrieving videos via extracted IDs and analyzing audio-visual content. For non-fluent languages, annotators used YouTube's auto-generated transcripts alongside visual elements. Advertisements were excluded if: (1) unavailable on YouTube at annotation time, (2) content was ambiguous despite translation, or (3) annotators lacked confidence in classification. This removed 189 advertisements (3.4%), leaving 5,422 labeled instances. Inter-annotator reliability reached Cohen's $\kappa = 0.79$ (substantial agreement), with all conflicts resolved through discussion.

Ethical Considerations. This study analyzes ads shown alongside publicly available "made-for-kids" YouTube videos using logged-out browsers. We did not interact with children and did not access, collect, or store any personal data. We collected only ads and advertiser metadata available via YouTube interfaces/APIs and the Ad Transparency Center, storing no viewer identifiers, cookies, or account information. Two adult annotators voluntarily labeled ads after training, with the option to skip disturbing content.

3 Results

In this section, we present the key findings from our data analysis. First, we detail the thematic analysis of our labeled ads, examining variations in the regional prevalence of inappropriate and child-directed ads as well as their content (Section 3.1). Next, we analyze advertiser-specific characteristics and their association with the dissemination of inappropriate ads across regions (Section 3.2).

3.1 Thematic Analysis

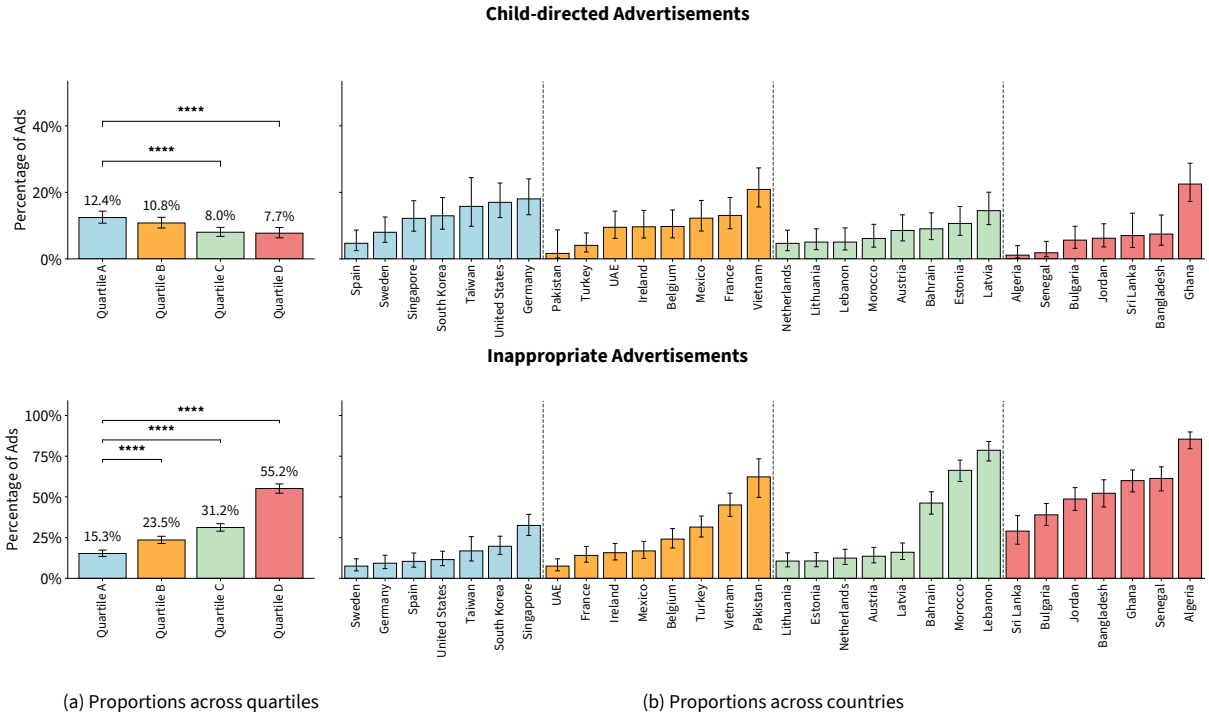
Following the annotation protocol described in Section 2.4, we compare the themes of 5,422 labeled ads across regions. Overall, 3236, 1659, and 527 ads were annotated as irrelevant, inappropriate, and child-directed, respectively.

Child-Directed Advertisements. Figure 1 shows child-directed ads consist of only a fraction of the ads displayed on child-oriented videos across the four quartiles (7.7-12.4%), with the majority of ads being labeled as irrelevant for children. This observation raises concerns about YouTube's ad delivery and pricing policies, as children may watch ads without realizing they are ads, something adults would typically skip, causing advertisers to pay for unintended engagement. Since YouTube charges advertisers only when an ad is viewed or engaged with [22], this issue may be amplified for children, who often struggle to distinguish ads from main video content [32]. The phenomenon worsens for countries in the lower quartile, with those in Quartile D being 1.6x less likely to feature child-directed ads, suggesting that advertising campaigns in countries with less stringent child safety policies may also face reduced efficacy in reaching child audiences. Interestingly, despite belonging to the lowest quartile, Ghana records the highest percentage of child-directed ads. This anomaly may be explained by Ghana's low ad pool diversity (Figure 9): only 12 unique child-directed ads account for 22.5% of all impressions, reflecting a reliance on a few ads for the majority of impressions.

Presence of Inappropriate Advertisements. Despite YouTube's efforts to ensure a safe viewing experience for children, we observe 30.6% of inappropriate ads in our sample. This concern deepens across quartiles, with countries in Quartile D showing approximately 3.6× more inappropriate ads than those in Quartile A (55.2% vs. 15.3%, $p < 0.001$, two-proportion z-test) (Figure 1). These results are consistent with the findings of Khan et al. [31], who previously observed an 8× difference (3.6% vs. 28.8%) in the proportion of inappropriate ads on child-directed videos between their analysis of 10 "high and low" policy countries. Notably, this disparity persists even two years after their study, and the overall proportion of inappropriate ads appears to have increased. European countries generally exhibit a lower percentage of inappropriate ads regardless of their COSI scores. For example, Lithuania, Estonia, the Netherlands, Austria, and Latvia each display only 10–16% inappropriate ads despite being in Quartile C, whereas Lebanon, Morocco, and Bahrain in the same quartile show much higher rates, ranging from 46–85%. This disparity may be partly explained by YouTube's enhanced ad transparency and accountability measures implemented in Europe following the Digital Services Act (2023), which requires platforms accessible to minors to protect the "mental and physical well-being" of children [9, 12, 20]. These measures include greater visibility into engagement statistics, complete date ranges of when ads were aired, and disclosures about ad removals, likely discouraging the circulation of inappropriate ads in European markets.

Our manual analysis of the inappropriate ads, as also confirmed by the video categories obtained from YouTube's API (Figure 2) shows that music videos constitute the majority of the inappropriate ads (50.9%), especially for quartiles C and D. We find music and entertainment-related videos to often contain themes such as suggestive or romantic content, inappropriate skin exposures, display of weapons, inappropriate language, drinking of alcohol, all which are categorically prohibited for child-oriented content as per Youtube's guidelines. Such findings highlight that YouTube could significantly mitigate the dissemination of inappropriate ads by using its video categories.

We also analyze the proportion of inappropriate ads across video language labels obtained from YouTube's API (Figure 3). Videos in certain languages, such as Punjabi, Hindi, or Arabic, generally contain a higher proportion of inappropriate ads (77.7-92.6%) compared to those in English, French, Spanish, or German (12.5-28.2%). While this observation may be partially explained by regional policy differences, as discussed earlier, the use of non-English languages may also pose challenges for YouTube's machine learning algorithms in accurately detecting and moderating inappropriate content due to their limited representation in training data [33, 39]. Moreover, ads without an assigned language label, accounting for 21.5% of our dataset, exhibited a substantially lower inappropriateness rate of 11.1%. Notably, 77.9% of these ads were associated with categories such as People & Blogs, Science & Technology, and Autos & Vehicles, which are less likely to host inappropriate content, whereas Music accounted for only 0.3% of such videos (Figure 2).



(a) Proportions across quartiles (b) Proportions across countries

Figure 1: Proportion of child-directed (top panel) and inappropriate (bottom panel) ads by quartiles (a) and country (b) based on their Child Online Safety Index (COSI) score. Error bars represent 95% confidence intervals. Asterisks indicate statistical significance determined by a two-proportion z-test (**$p < 0.001$) relative to quartile A.**

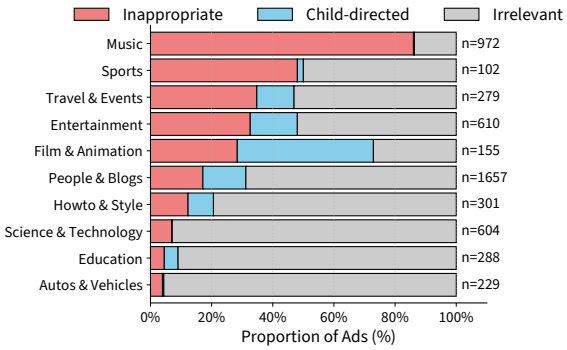


Figure 2: Proportion of ad content by category. Each bar represents 100% of the ads within a given category, segmented as irrelevant, child-directed, or inappropriate. The total sample size (n) is annotated.

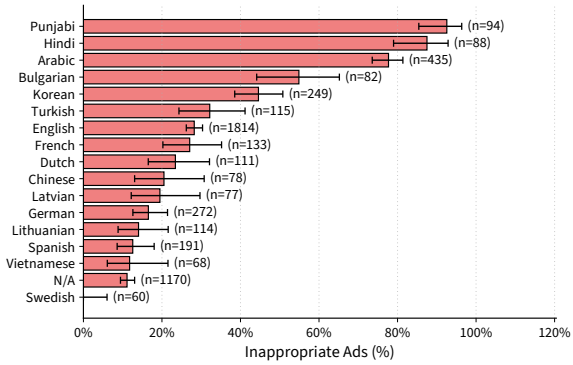


Figure 3: Proportion of inappropriate ads by language. The total sample size (n) for each language is annotated on the right of the bar.

Overall, the findings highlight questions about YouTube's lack of consistency in moderating ad-content across regions, with such significant regional differences suggesting a potential selective prioritization of audiences' experiences. While regional ad preferences or variations in ad-auctioning algorithms may partly explain this phenomenon, the findings raise serious safety concerns. Repeated exposure to inappropriate material, particularly when embedded within content that children may perceive as child-directed, risks desensitizing young audiences to such themes and shaping their attitudes and purchasing behaviors toward inappropriate products. This is especially problematic given evidence that children struggle to distinguish advertising from the main content [32].

Key Insights

- Up to 20.6% of advertisements across quartiles are served by unverified advertisers.
- Unverified advertisers are more likely to serve inappropriate advertisements (35.3% vs 29.9%, $p < 0.01$)

- Unverified advertisers tend to serve inappropriate advertisements in lower quartiles more often, indicating a presence of lower-quality advertisers in the markets.
- The presence of native advertisements decreases starkly across the quartiles (69.3% to 22.8%), risking the sponsorship of culturally-sensitive products/themes.
- A significant proportion of inappropriate ads tends to come from advertisers based in Ecuador and Lebanon, where 90%+ of ads from both countries are inappropriate.

3.2 Advertiser Analysis

Next, we analyze advertiser characteristics and their association with inappropriate ad delivery. Google introduced its identity verification program in 2018 for political advertisers, later expanding it to all advertiser categories in 2020 [8, 21, 25, 34]. The program mandates verification of business identity, operational jurisdiction, and promoted products or services. Since 2022, Google discloses advertiser name, identity, and verification status via the "My Ad Center" interface. However, advertisers may continue serving ads either until selected for verification or for 30 days post-notification; non-compliance results in account suspension [21]. Advertisers in sensitive sectors (e.g., finance) and those promoting high-risk content receive prioritized verification. In this study, we used Google's disclosure features to collect advertiser metadata via programmatic queries to the "My Ad Center" during ad collection using logged-out accounts. For each impression, we recorded advertiser name, declared location, and verification status. Across 22,766 ad impressions, we identified 2,928 unique advertisers (deduplicated by name) distributed across 101 countries.

Advertiser Verification. Our analysis shows that unverified advertisers account for 20.6%, 13.0%, 6.9%, and 16.1% of overall ads in Quartiles A, B, C, and D, respectively (Figure 4a), with statistically significant inter-quartile differences ($p < 0.01$, two-proportion z-test). Quartile A exhibits the highest unverified rate despite its high COSI, while Quartile C, dominated by European countries, records the lowest rate, consistent with stricter regional enforcement (Figure 10). Unverified advertisers deliver *inappropriate ads* at significantly higher rates (35.3%) than verified advertisers (29.9%, $p < 0.01$, two-proportion z-test). This disparity manifests differently across quartiles (Figure 4b): Quartile A shows equivalent inappropriate ad rates for both unverified and verified advertisers (~15% each, $p = 0.47$), while Quartiles B, C, and D exhibit significant gaps ($p < 0.05$, $p < 0.05$, $p < 0.01$ respectively). In absolute terms, inappropriate ad rates from unverified advertisers escalate from 15% (Quartile A) to 30% (B), 47% (C), and 67% (D), representing 3.1× and 4.5× increases in Quartiles C and D relative to Quartile A. These patterns indicate that verification status correlates with ad safety only in lower COSI regions. The quartile-specific effects suggest differential enforcement rigor or advertiser quality distributions across markets. Critically, verified advertisers still deliver 29.9% inappropriate ads overall (54% in Quartile D), indicating systematic verification process failures.

Advertiser Location. Next, we analyze the reported geographical locations of advertisers. Figure 5 shows a monotonic decrease in native advertiser proportions across quartiles: 69.3% (Quartile A) to 22.8% (Quartile D). Non-native advertisements exhibit significantly higher rates of child-inappropriate content (36.9% vs. 21.5%, two-proportion z-test, $p < 0.0001$). This disparity may reflect advertiser resource asymmetries in digital ad auctions, where higher-quartile countries dominate bidding. Consequently, countries with limited native advertising face disproportionate exposure to content misaligned with local culture or social norms. For example, Bahrain received alcohol-themed music videos from Ecuadorian advertisers and sexually suggestive entertainment content from Italian and Australian advertisers. Bangladesh similarly received advertisements featuring revealing attire. Although YouTube's global policies flag such content as child-inappropriate, these placements may affect region-specific cultural sensitivities.

Figure 6 identifies Lebanon and Ecuador as dominant sources of inappropriate ads to Quartile C and D countries, accounting for 90.6% and 93.5% of country-specific totals, respectively, across 120 distinct advertisers. Ecuador's prominence is anomalous given the absence of South American countries in our dataset and lack of cultural or linguistic ties to targeted Arab and South Asian regions. Music content comprises 94.8% of inappropriate ads from both sources. Language metadata reveals systematic mismatches: Lebanese ads predominantly use Arabic (87.4%, regionally appropriate), whereas Ecuadorian ads feature Indian music videos with Indian language codes (51.9%: Punjabi, Hindi, Odia, Telugu, Malayalam, Marathi, Kannada, Tamil), English (28.5%), and Arabic (12.0%). Google-verified advertisers account for 84.8% of these inappropriate placements. U.S. and South Korean advertisers similarly target Arab, South Asian, and African markets with inappropriate content, facilitated by higher advertising budgets [42]. Among high-transparency countries (Quartile A), Türkiye represents the largest source (31.4%) of inappropriate cross-border advertising. These findings reveal fundamental verification framework failures. Google-verified advertisers systematically place culturally misaligned content across geographic and linguistic boundaries Current verification mechanisms inadequately account for cross-cultural appropriateness. These patterns suggest that incorporating advertiser location more directly into content moderation and placement decisions could strengthen protections against inappropriate advertising.

Advertiser Analysis. We conducted a retrospective verification analysis of the 262 inappropriate ads served by unverified advertisers in our dataset. We performed this analysis 45-60 days after ad collection, exceeding Google's 30-day advertiser verification window [18, 25]. Our analysis relies on Google's Ad Transparency Center, which provides data on advertisers across Google's platforms. According to Google, advertiser details may be missing if they have not completed verification, and ads from unverified advertisers may only be visible if served in Europe or Türkiye [17, 20]. The platform supports advertiser search with ad pool visualization.

For each of the 262 ads, we queried the Ad Transparency Center using the advertiser name and serving region, filtering for YouTube video ads. When multiple advertisers shared identical names and locations, we manually inspected all associated ad creatives to establish ground-truth matches and extract verification status. Following our matching protocol (Appendix 7.2), we identified 166 ads (63.4%) served by advertisers that remained unverified beyond the 30-day deadline. This demonstrates that a substantial fraction of advertisers serve inappropriate content without completing verification.

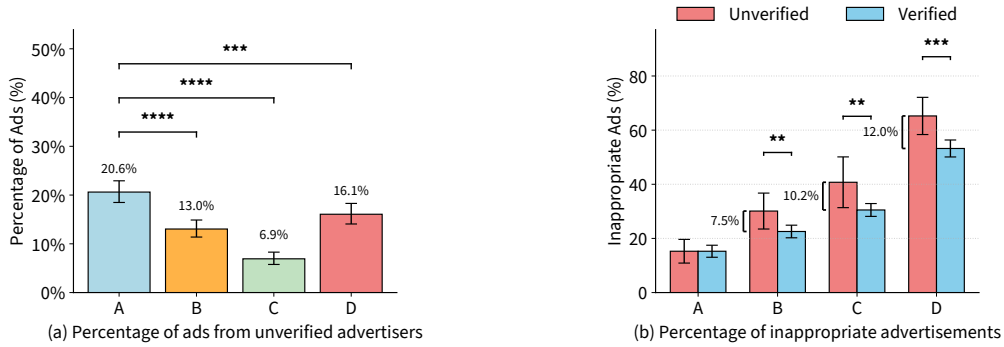


Figure 4: Percentage of ads served by unverified advertisers (a) and percentage of inappropriate ads by verified and unverified advertisers across COSI quartiles. Error bars represent 95% confidence intervals. Asterisks indicate statistical significance determined by a two-proportion z-test ($^{*}p < 0.1$,$^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$). In (a), all quartiles are compared to Quartile A. In (b), significance is shown for the difference between unverified and verified advertisers within each quartile.

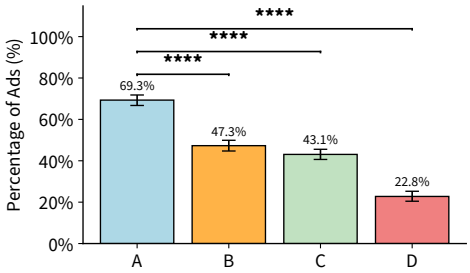


Figure 5: Percentage of ads across COSI quartiles served by advertisers native to the country of placement. Error bars represent 95% confidence intervals. Asterisks indicate statistical significance determined by a two-proportion z-test ($^{****}p < 0.001$) relative to quartile A.

These findings, combined with our detection results, indicate that unverified advertisers systematically exploit the verification grace period to disseminate policy-violating ads before enforcement. We further identified 12 cases where advertisers eventually achieved verification, 9 Ecuador-based entities serving inappropriate music videos to audiences in Algeria, Bangladesh, Pakistan, and Türkiye. This pattern highlights potential weaknesses in YouTube's verification system and points to the need for greater transparency regarding advertisers' intended audiences and the types of content they promote.

3.3 Automated Detection Analysis

In Section 3.1 and 3.2 we analyzed ad metadata including categories, language, location data, and advertiser attributes. A key question is whether these readily available metadata features alone can enable effective automated detection. We demonstrate that even lightweight machine learning models trained exclusively on ad metadata achieve high classification accuracy, suggesting that metadata-based approaches can complement or reduce reliance on computationally expensive content analysis.

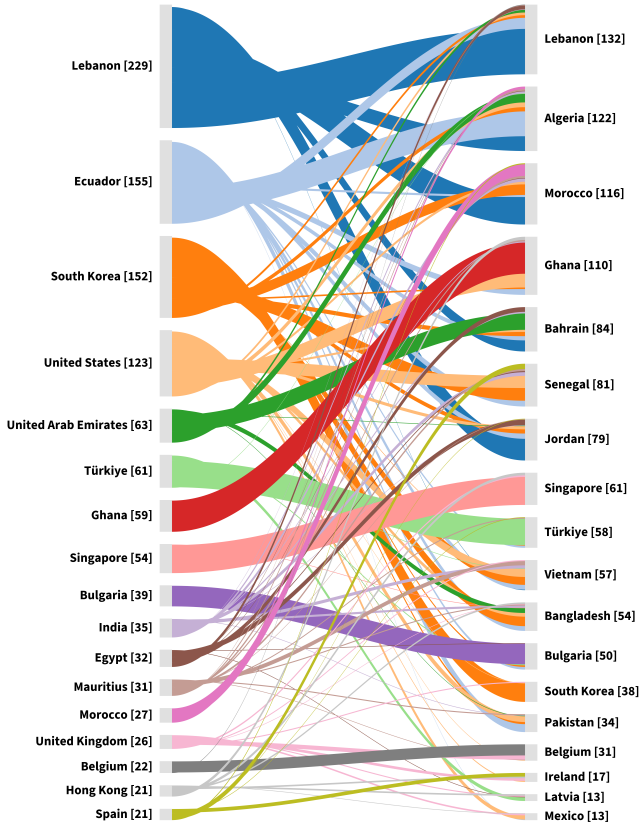


Figure 6: Inappropriate ad targeting relationships between advertiser's location (left) and ad placement (right). The plot has been limited to include advertiser locations placing at least 20 ads, and countries with at least 30 total inappropriate ad placements. The counts represent the number of ads advertised from (left) or served in a country (right)

Features and Model. Our dataset contains categorical variables such as video location, advertiser location, and language. We derived additional features from these, including country and language groupings. To avoid extreme feature sparsity from one-hot

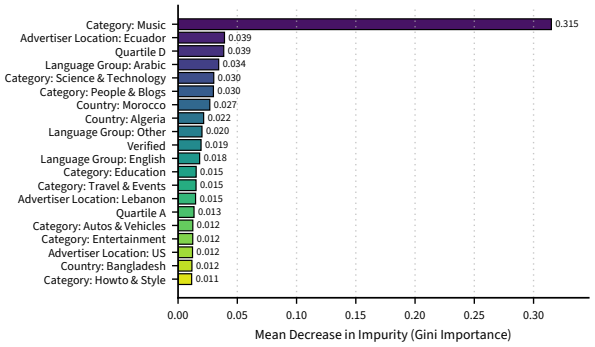


Figure 7: Top 20 Features from Random Forests Classifier

encoding, highly sparse features (>99%) were removed, producing more generalizable representations. For model selection, we used an 80-20 train-test split and stratified 5-fold cross-validation to account for class imbalance. Multiple classifiers were evaluated, with Random Forest, Logistic Regression, and CatBoost achieving the best performance and selected for hyperparameter tuning. Random forest emerged as the top performer with 83.83% accuracy and 75.13% unweighted F1-score. Importantly, this accuracy is obtained only using the features analyzed earlier, other features outside the scope of this paper like engagement metrics were not considered. Further details on the model selection methodology and hyperparameter tuning results are provided in the appendix.

Feature Importance. Figure 7 displays the top 20 most influential features for classification with our tuned Random Forests model. Complementing our findings, we observe that music category dominates. This is consistent with our earlier finding that music videos constitute 50.9% of inappropriate ads. Other influential features also contribute notably, including language (Arabic and Punjabi), advertiser verification status, location (Equador, Lebanon, and the US), and COSI quartiles (A and D), which further highlights the feasibility of leveraging these attributes to more effectively filter inappropriate content for children.

4 Discussion

Need for Additional Transparency Measures. Countries in lower COSI quartiles face up to 3.6× higher exposure to inappropriate ads than those in higher quartiles. Europe is a notable outlier with lower rates and higher advertiser verification relative to COSI levels. After the EU's 2023 Digital Services Act, YouTube added advertiser and ad-level disclosures in the EEA and Türkiye (impression ranges, dates, audience criteria, format/topic, Google Ad Grants status) and began reporting removals with links to the Ads Transparency Center [12, 24]. Extending comparable transparency and removal reporting globally could improve accountability and narrow safety gaps.

Gaps in Google's Advertiser Verification Program. Unverified advertisers are more likely to serve inappropriate ads, especially in lower COSI quartiles. The current grace period (up to 30 days) allows ad delivery before verification, and many advertisers never complete verification, enabling repeat abuse via new accounts. YouTube should prevent unverified advertisers from targeting child-oriented content and enforce tighter deadlines. Because some inappropriate ads also come from verified advertisers, verification should include stronger legal attestations about intended audience and content.

Leveraging Feature Signals to Improve Relevance and Safety. We find that 59.8% of ads are irrelevant for children. Cross-border targeting often compounds this problem; for example, South Korean advertisers reached South Asia, Arab, and African audiences even though 76% of their ads were labeled in Korean (Figure 6). Language labels, advertiser location, content category, verification status, video duration, and placement are useful signals for filtering irrelevant and risky ads. Prioritizing these signals in countries with lower COSI scores can reduce exposure disparities.

5 Related Works

Detecting unsafe child content (video-level). Prior work classifies inappropriate or unsafe material in child-directed videos using metadata, subtitles, and audio-visual cues, reporting strong accuracy on large corpora [2, 4, 30, 44, 48, 50]. These studies focus on video content itself rather than ads or advertiser attributes.

Advertising on child content (exposure and marketing). Studies measuring ads alongside children's videos document high ad load, persuasive tactics, and substantial exposure to branded content; Liu et al. report that 26.9% of child-oriented videos include at least one inappropriate ad [15, 16, 35, 53]. However, this line of work rarely examines regional disparities or links exposure to advertiser characteristics.

Enforcement, policy, and advertiser attributes. Khan et al. show large cross-region differences in ad-duration policy violations across 10 regions [31]. We extend to 32 regions and use COSI quartiles to analyze disparities, adding advertiser-level factors (verification status, location) enabled by the 2023 Ads Transparency Center [17]. Related work on personalization (e.g., Mai et al.) examines targeting with logged-in profiles but not advertiser verification or location in relation to inappropriate content [37]. To our knowledge, this is the first systematic study connecting advertiser verification and geography to the prevalence of inappropriate ads across regions.

6 Conclusion

Both advertisers and policy strength shape children's exposure to inappropriate YouTube ads, but our large cross-regional audit underscores the advertiser lever as especially actionable and underexplored. Weak or absent verification and opaque cross-border provenance consistently elevates risk, particularly in weaker-policy regions. While prior work has linked policy to safer outcomes (often at smaller scales), our results show that tightening advertiser verification and provenance transparency can materially reduce harm alongside policy.

References

[1] Adzoola. 2024. *YouTube Ads Benchmark & Insights 2024–2025.* https://www.adzoola.com/youtube-ads-benchmark-insights-24-25/

[2] Saeed Ibrahim Alqahtani, Wael M. S. Yafooz, Abdullah Alsaeedi, Liyakathunisa Syed, and Reyadh Alluhaibi. 2023. Children's Safety on YouTube: A Systematic Review. 13, 6 (2023), 4044. doi:10.3390/app13064044

[3] Anonymous. 2025. Advertisers, Provenance, and Policy: Repository. https://anonymous.4open.science/r/Advertisers-Policy-and-Child-Safety-8BB5. accessed: 2025-10-08.

[4] Le Binh, Rajat Tandon, Chingis Oinar, Jeffrey Liu, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Fred Morstatter, Simon Woo, and Jelena Mirkovic. 2022. Samba: Identifying Inappropriate Videos for Young Children on YouTube. 10 pages. https://doi.org/10.1145/3511808.3557442

[5] BizAsiaLive. 2025. Youngsters continue to turn away from traditional TV; radio listening up. https://www.bizasialive.com/youngsters-continue-watching-traditional-tv-less-frequently-radio-listening-up/.

[6] Social Blade. 2025. https://socialblade.com/

[7] Angela J. Campbell. 2016. Rethinking Children's Advertising Policies for the Digital Age. *Loyola Consumer Law Review* 29, 1 (2016), 1–52.

[8] John Canfield. 2020. Increasing transparency through advertiser identity verification. *Google* (June 2020). https://blog.google/products/ads/advertiser-identity-verification-for-transparency

[9] Foo Yun Chee. 2023. *Google vows more transparency on ads as new EU rules kick in.* https://www.reuters.com/technology/google-vows-more-transparency-ads-new-eu-rules-kick-2023-08-24/

[10] Children and Screens. 2020. Advertising and Kids. https://www.childrenandscreens.org/learn-explore/research/advertising-and-kids/.

[11] DataReportal. 2025. Essential YouTube Statistics and Trends for 2025. https://datareportal.com/essential-youtube-stats

[12] European Commission. 2025. The impact of the Digital Services Act on digital platforms. https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms.

[13] European Union. 2016. General Data Protection Regulation (GDPR) — Legal Text. https://gdpr-info.eu/

[14] Federal Trade Commission. 2025. Children's Online Privacy Protection Rule (COPPA). https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa

[15] Frances Fleming-Milici, Haley Gershman, Hanako O. Agresta, Melissa McCann, and Jennifer Harris. 2025. Young Children's (Aged 3 to 8 Years) Food and Beverage Brand Exposure on YouTube and YouTube Kids: An Observational Study and Content Analysis. 125, 10 (2025), 1482–1493.e2. doi:10.1016/j.jand.2025.05.010

[16] Jason Freeman, Jeff Conlin, Christina Triptow, Michael Burke, and Jin Chen. 2025. Promoting with Kid Gloves: Analysis of YouTube "Kidfluencer" Videos Before and After Settlement for COPPA Violations. *Journal of Current Issues & Research in Advertising* 0, 0 (2025), 1–19. doi:10.1080/10641734.2025.2547276

[17] Google. 2025. https://adstransparency.google.com/

[18] Google. 2025. About Google Ads account pausing. https://support.google.com/adspolicy/answer/9872152.

[19] Google. 2025. Ad-serving protections for children. https://support.google.com/adspolicy/answer/14170968. Accessed: 2025-09-29.

[20] Google. 2025. Ads Transparency Center FAQ. https://adstransparency.google.com/faq.

[21] Google. 2025. Advertiser verification - Advertising Policies Help. https://support.google.com/adspolicy/answer/9703665?hl=en

[22] Google. 2025. Grow your business with Google Ads - Google Ads Help. https://support.google.com/google-ads/answer/6336021?hl=en

[23] Google. 2025. Youtube Data API v3 Documentation. https://developers.google.com/youtube/v3/docs

[24] Google Cloud. 2025. Google Ads Transparency Center — BigQuery Public Data. https://console.cloud.google.com/marketplace/details/bigquery-public-data/google-ads-transparency-center?project=gen-lang-client-0152968505

[25] Megan Graham. 2020. Google will make all advertisers prove their identities, so people can see who they are and which country they're in. https://www.cnbc.com/2020/04/23/google-advertiser-verification-process-now-required.html#:~:text=Google%20said%20it%20previously%20had,by%20humans%2C%20a%20spokeswoman%20said.

[26] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research* 15, 9 (Oct. 2005), 1277–1288. doi:10.1177/1049732305276687

[27] L. Rowell Huesmann. 2007. The Impact of Electronic Media Violence: Scientific Theory and Research. *Journal of Adolescent Health* 41, 6 Suppl 1 (2007), S6–S13. doi:10.1016/j.jadohealth.2007.09.005

[28] ICUC. 2025. YouTube Content Moderation: A Comprehensive Overview. https://icuc.social/resources/blog/youtube-content-moderation/

[29] DQ Institute. 2023. Child Online Safety Index 2023. https://www.dqinstitute.org/impact-measure/

[30] Rishabh Kaushal, Srishty Saha, Payal Bajaj, and Ponnurangam Kumaraguru. 2016. KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube. 157–164 pages. doi:10.1109/PST.2016.7906950

[31] Emaan Bilal Khan, Nida Tanveer, Aima Shahid, Mohammad Jaffer Iqbal, Haashim Ali Mirza, Armish Javed, Ihsan Ayyub Qazi, and Zafar Ayyub Qazi. 2024. Analyzing Ad Exposure and Content in Child-Oriented Videos on YouTube. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1215–1226. doi:10.1145/3589334.3645585

[32] David Kirkpatrick. 2016. Study: 82% of middle-school kids can't distinguish ads from actual news. *Marketing Dive* (Nov. 2016). https://www.marketingdive.com/news/study-82-of-middle-school-kids-cant-distinguish-ads-from-actual-news/430934/

[33] Yi-Ting Lien and Maya Vishwanath. 2025. Linguistic Inequity in Facebook Content Moderation. *Technology Science* (2025). https://techscience.org/a/2025022501/ Published February 25, 2025; accessed 2025-10-02.

[34] Andrew Liptak. 2018. Google says political-leaning advertisers will require an ID to verify their identity. *The Verge* (May 2018). https://www.theverge.com/2018/5/5/17322744/google-political-election-ads-verification-advertisers

[35] Jeffrey Liu, Rajat Tandon, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Neelesh Gupta, Zoe Zhou, and Jelena Mirkovic. 2022. *Did Your Child Get Disturbed by an Inappropriate Advertisement on YouTube?* doi:10.48550/ARXIV.2211.02356

[36] Renkai Ma and Yubo Kou. 2022. A Study of Fairness Perception of Content Moderation on YouTube *(CSCW '22, Vol. 6)*. Article 425. doi:10.1145/3555150

[37] Cat Mai, Bruno Coelho, Julia Kieserman, Lexie Matsumoto, Kyle Spinelli, Eric Yang, Athanasios Andreou, Rachel Greenstadt, Tobias Lauinger, and Damon McCoy. 2025. More and Scammier Ads: The Perils of YouTube's Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2025, 4 (2025), 1014–1038. doi:10.56553/popets-2025-0169

[38] Meta Platforms, Inc. 2025. Online Child Protection. https://www.meta.com/safety/topics/online-child-protection

[39] Gabriel Nicholas and Arjun Bhatia. 2023. Toward Better Automated Content Moderation in Low-Resource Languages. *Journal of Online Trust and Safety* 2, 1 (2023). doi:10.54501/jots.v2i1.15

[40] NordVPN. 2025. *Best VPN Server Locations: View the Full List.* https://nordvpn.com/servers/?srsltid=AfmBOorPzLSLHuc4KiGWoNvV-Kajtem8vub-0hpTp5EVE5S_o3FOiS9Y

[41] Carly Nyst. 2019. *Children and Digital Marketing: Rights, Risks and Opportunities.* Technical Report. United Nations Children's Fund (UNICEF).

[42] Oberlo. 2025. https://www.oberlo.com/statistics/digital-ad-spend-by-country

[43] Jessica Packer, Helen Croker, Anne-Lise Goddings, Emma J Boyland, Claire Stansfield, Simon J Russell, and Russell M Viner. 2022. Advertising and Young People's Critical Reasoning Abilities: Systematic Review and Meta-analysis. *Pediatrics* 150, 6 (2022), e2022057780. doi:10.1542/peds.2022-057780

[44] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2021. *Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children.* arXiv:1901.07046 [cs] doi:10.48550/arXiv.1901.07046

[45] Prithivi Raj and Lalit Deb. 2025. Cross-Border Comparative Advertising: Legal Framework and International Perspectives. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5135969

[46] World Population Review. 2025. Youtube Penetration Index. https://worldpopulationreview.com/country-rankings/youtube-users-by-country

[47] Christianna Silva. 2025. YouTube is relying more on human content moderators. https://mashable.com/article/youtube-human-content-moderation

[48] Shubham Singh, Rishabh Kaushal, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2019. KidsGUARD: fine grained approach for child unsafe video representation and detection. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. Association for Computing Machinery, New York, NY, USA, 2104–2111. doi:10.1145/3297280.3297487

[49] Statisa. 2025. https://www.statista.com/topics/7666/internet-advertising-worldwide/?srsltid=AfmBOoqeS4gQteWxmZI8hyMF1EMddJcNK2nyZWJfJhpDKJOccg8YqLg_#topicOverview

[50] Rashid Tahir, Faizan Ahmed, Hammas Saeed, Shiza Ali, Fareed Zaffar, and Christo Wilson. 2020. Bringing the kid back into YouTube kids: detecting inappropriate content on video streaming platforms. 6 pages. https://doi.org/10.1145/3341161.3342913

[51] Global Media Insight Team. 2025. YouTube Statistics 2025 (Demographics, Users by Country & More). https://www.globalmediainsight.com/blog/youtube-users-statistics/.

[52] Paul J. Wright, Bryant Paul, and Debby Herbenick. 2021. Preliminary Insights from a U.S. Probability Sample on Adolescents' Pornography Exposure, Media Psychology, and Sexual Aggression. *Journal of Health Communication* 26, 1 (2021), 39–46. doi:10.1080/10810730.2021.1887980

[53] Samantha L. Yeo, Alexandria Schaller, Michael B. Robb, and et al. 2021. Frequency and Duration of Advertising on Popular Child-Directed Channels on a Video-Sharing Platform. *JAMA Network Open* 4, 5 (2021), e219890. doi:10.1001/jamanetworkopen.2021.9890

[54] Youtube. 2025. Youtube Monetization Markets 2025. https://support.google.com/youtube/answer/1342206?hl=en

[55] YouTube Help Center. 2025. Advertiser-friendly content guidelines. https://support.google.com/youtube/answer/6162278?hl=en#zippy=%2Cpolicy-detail

7 Appendix

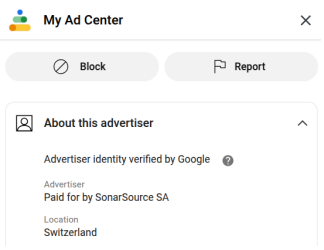


Figure 8: Example of advertiser detail disclosures shown in the My Ad Center pop-up for advertisements.

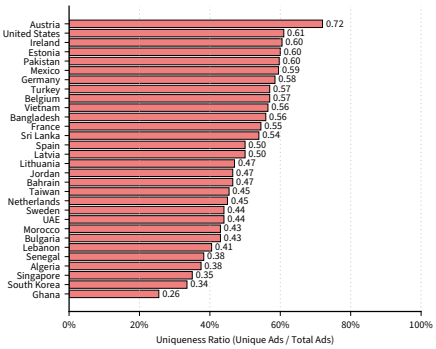


Figure 9: Ad uniqueness ratio (unique ad IDs / total ads collected) per country.

7.1 Restoring Advertiser Details

Advertiser details in our dataset were incomplete for 5,946 ad impressions in our collected dataset of 26,737 ads. We restored advertiser details for 3,793 impressions by using information from duplicate ads sharing the same `Ad ID` within the same country. This approach was reasonable, as our validation revealed that inconsistencies were rare: among 7,550 unique `(Country, Ad ID)` pairs in our dataset, only 116 (1.54%) showed conflicting advertiser names. This indicates the number of impressions with potentially misattributed advertiser details is expected to be less than 57 (only 0.25% of our dataset), a trivial figure that does not impact our subsequent analysis. The remaining 2,153 impressions were removed from analysis as explained in Section 2.3.

7.2 Post-hoc analysis details

Table 2: Post-hoc analysis of Advertiser Verification Status

Category	Count	Status
Total ads analyzed	260	–
Advertiser not located	121	Unverified
Advertiser located, ad found	57	–
- Remained unverified	45	Unverified
- Completed verification	12	Verified
Advertiser located, ad not found	82	Uncertain

From our labeled dataset, we analyzed all 262 inappropriate ads placed by unverified advertisers. For each ad, we searched the associated advertiser on Ad Transparency Center by name and specified the region in which the ad was shown, filtering for video ads on YouTube. In cases where multiple advertisers shared the same name and reported location, we manually inspected all associated ads to identify the matching ads and their advertisers to record their verification status. Two ads were excluded from the analysis because their associated advertiser had over 50,000 ads, making it infeasible to manually locate the specific ads among such a large set. We were unable to locate the associated advertisers for 121 of the remaining ads, likely because they remain unverified or were removed from the Transparency Center [20]. Among the 139 advertisers we did locate, we found the specific ad in 57 cases: 45 advertisers were still unverified, while 12 had since completed verification. For the remaining 82 cases, we located the advertiser but not the observed ad, which may indicate that the ad was removed or was not displayed in the Transparency Center because the advertiser was unverified and serving ads outside Europe. Overall, this suggested that at least 166 (121+45) ads appeared to be associated with advertisers that tend to remain unverified well beyond the general deadline, while at least 12 of the associated advertisers were eventually verified. Table 2 summarizes our findings.

7.3 Automated Detection Details

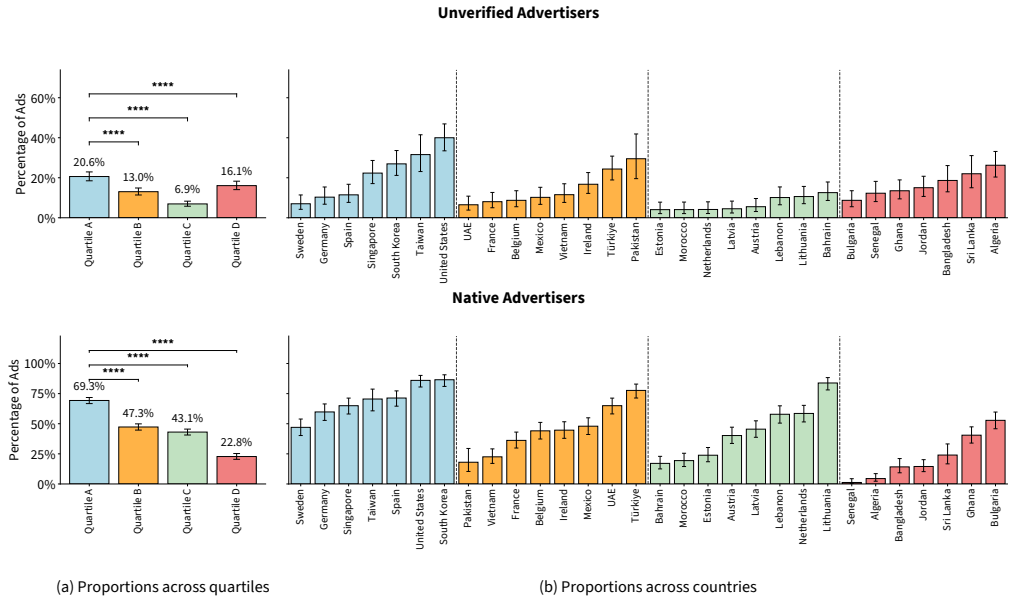
Table 3: Cross-Validation Accuracy of Baseline Models (Mean ± SD)

Classifier	Accuracy	Classifier	Accuracy
Random Forest	**0.84 ± 0.01**	KNN	0.83 ± 0.01
ExtraTrees	0.81 ± 0.003	**Logistic Reg.**	**0.85 ± 0.01**
CatBoost	**0.84 ± 0.01**	MLP	0.80 ± 0.01
Naive Bayes	0.62 ± 0.02	LightGBM	0.81 ± 0.01
XGBoost	0.81 ± 0.01	HistGB	0.81 ± 0.01

Table 4: Train (CV) and Test Set Performance Metrics for Top Models

Classifier	CV Accuracy	Test Set			
		Accuracy	Avg F1	Avg Precision	Avg Recall
Random Forest (RF)	0.843	**0.838**	**0.75**	0.83	**0.72**
Logistic Reg. (LR)	**0.846**	0.832	0.74	0.82	**0.72**
CatBoost (CB)	0.842	0.836	0.74	**0.85**	0.71

We first evaluated multiple classification methods, with average results shown in Table 3. Random Forest (RF), Logistic Regression (LR), and CatBoost (CB) achieved the top three cross-validation accuracies and were selected for hyperparameter tuning. We performed exhaustive grid searches with 180, 40, and 270 configurations for RF, LR, and CB respectively. Table 4 presents the final results with optimal hyperparameters. All three models demonstrated robust performance, with Random Forests emerging as the best performer with 83.8% test accuracy.



(a) Proportions across quartiles (b) Proportions across countries

Figure 10: Percentage of ads served by unverified advertisers (top panel) and native advertisers (bottom panel) across (a) quartiles and (b) countries. Error bars represent 95% CIs. Asterisks indicate statistical significance determined by a two-proportion z-test ($^{*}p$<0.1, $^{**}p$ < 0.05, $^{***}p$ < 0.01, $^{****}p$ < 0.001) relative to quartile A.

Table 5: Ads collected, sampled, and non-ambiguous counts across countries.

Country	Collected	Sampled	Non-ambiguous
Algeria	286	200	179
Austria	2066	200	199
Bahrain	578	200	199
Belgium	1245	200	195
Bulgaria	889	200	195
Estonia	1719	200	197
Ireland	2559	200	197
France	645	200	199
Germany	1034	200	194
Ghana	277	200	200
Latvia	1199	200	200
Jordan	583	200	193
Lebanon	646	200	178
Lithuania	671	200	198
United States	1450	200	200
Mexico	1076	200	196
Morocco	265	200	196
Netherlands	422	200	193
South Korea	658	200	193
Singapore	389	200	197
Sweden	899	200	200
Spain	312	200	192
UAE	801	200	200
Vietnam	369	200	182
Turkey	1117	200	197
Senegal	196	196	163
Bangladesh	136	136	134
Taiwan	110	110	95
Sri Lanka	102	102	100
Pakistan	67	67	61

Table 6: Codebook for child-directed content categories

Category	Description
Cartoons	Animated shows or movies aimed toward children.
Storytelling	Videos with storytelling elements (non-animated); may include children's shows, vlogs, etc.
Books and Literature	Videos about reading books, comics, or magazines for children.
Family friendly gaming	Videos featuring video games designed for children.
DIY & Arts and Crafts	Videos about arts and crafts, including instructions or demonstrations.
Toys	Videos featuring promotions, reviews, unboxing, demonstrations, or play sessions with toys for kids.
Educational Content	Educational content made for children (e.g., science experiment demos, videos promoting learning apps and games).
Nursery Rhymes & Music	Rhymes, songs, poems, or musical compositions intended for children.
Play and Adventure	Videos emphasising physical and adventure activities for children (e.g., dancing, rock-climbing, theme parks, camping).
Health & Hygiene	Health and wellness videos aimed at children (e.g., kids' soap, toothpaste, shampoo).
Kid's Fashion	Promotions of kids' fashion items such as clothing, school bags, shoes, and accessories.
Movies	Children's movies, including animated and family-friendly films (rated suitable for children under 13).
Cooking and Food	Videos involving food items or cooking demonstrations targeting children.

Table 7: Codebook of inappropriate content categories and their definitions.

Category	Definition / Examples
18+ Interests	• *Alcohol*: Products such as alcohol, including vineyard tours. • *Tobacco*: Tobacco products and related items (e-cigarettes). • *Recreational Drugs*: Recreational drugs and related paraphernalia. • *Gambling and Casino Games*: Online or real-world gambling, lotteries, betting, casino games.
Age-sensitive Media	• *Teen and Adult Media*: Ads for movies, TV shows, or console games not suitable for children.
Dangerous Products or Services	• *Dangerous Content*: Activities or products that require adult supervision (e.g., paintball, airsoft, ax-throwing, fireworks, weapons).
Financial	• *Complex Speculative Financial Products*: Contracts for difference, rolling spot forex, financial spread betting.
Health and Beauty	• *Beauty and Cosmetics*: Cosmetics and personal care focused on body image. • *Body Modification and Weight Loss*: Cosmetic procedures, weight loss, tanning, piercings, tattoos. • *Food and Beverage*: Consumable food and drinks. • *Health and Wellness*: Healthcare, reproductive health, substance-abuse recovery, eating disorders, health insurance. • *Pharmaceuticals and Supplements*: Medications, vitamins, and supplements.
Privacy, Safety, and Gimmicks	• *Contests and Sweepstakes*: Promotional contests or sweepstakes. • *Mobile Subscriptions*: Services requiring a mobile number. • *Social Networks*: Platforms that connect users. • *Spying and Arrest Records*: Services implying spying or providing arrest records. • *Quizzes*: Personality quizzes requiring personal data. • *Video Game Skins / Loot Boxes*: Selling or trading skins, loot boxes. • *Virtual Worlds / Adult Chat Rooms*: Platforms for adults to meet/connect with strangers.
Sensitive or Controversial	• *Astrology, Occult, Paranormal*: Astrology, occult, paranormal content. • *Politics, Religion, Sensitive Social Issues*: Ads about political, religious or controversial social topics.
Sexual and Romantic	• *Adult / Sexually Suggestive Content*: Sexual or mature content intended for adults. • *Dating and Relationships*: Dating services, matchmakers, relationship advice. • *Romantic Content*: Romance genre materials (games, books, comics). • *Significant Skin Exposure*: Revealing clothing or suggestive imagery (e.g., swimwear, underwear modeling).
Violent, Scary, or Crude	• *Fight Sports and Martial Arts*: Boxing, wrestling, martial arts, self-defense. • *First-Person Shooters / Battle Games*: Games involving shooting or army combat. • *Shocking or Scary Content*: Violent, gruesome, graphic or profane content (e.g., zombies, scary imagery).