Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Inconsistencies in Classification of Online News Articles: A Call for Common Standards in Brand Safety Services

Michael E Smith[1], Riley Grossman[1], Antonio Torres-Agüero[2], Pritam Sen[1], Cristian Borcea[1], Yi Chen[1]

[1]New Jersey Institute of Technology, Newark, New Jersey, USA
[2]DeepSee.io, Los Angeles, California, USA
{mes6, rag24, ps37, borcea, yi.chen}@njit.edu, antonio@deepsee.io

Abstract

This study examines inconsistencies in the brand safety classifications of online news articles by analyzing ratings from three leading brand safety providers, DoubleVerify, Integral Ad Science, and Oracle. We focus on news content because of its central role in public discourse and the significant financial consequences of unsafe classifications in a sector that is already underserved by digital ad spending. By collecting data from 4,352 news articles on 51 domains, our analysis shows that brand safety services often produce conflicting classifications, with significant discrepancies between providers. These inconsistencies can have harmful consequences for both advertisers and publishers, leading to misplaced advertising spending and revenue losses. This research provides critical insights into the shortcomings of the current brand safety landscape. We argue for a standardized and transparent brand safety system to mitigate the harmful effects of the current system on the digital advertising ecosystem.

Data — https://github.com/rag24/brand_safety_tool_audit

1 Introduction

Brand safety services are essential for advertisers to ensure that their ads do not appear alongside harmful or inappropriate content. Their value is best demonstrated in a recent paper that surveyed over 200 decision makers from companies spending more than $10 million annually on advertising, finding that more than 80% were concerned with brand safety. These same participants indicated that they would pay a nearly 50% premium to guarantee that their advertisements would only appear alongside safe content (Johnson, Voorhees, and Khodakarami 2023), thus highlighting the need for brand safety classification tools.

Our study focuses on DoubleVerify (DV), Integral Ad Science (IAS) and Oracle, because they are three of the most widely used state-of-the-art providers of brand safety technology in the industry (Elmore 2024; Business Insider 2024). Each brand safety provider offers a different, proprietary classification system to assess the safety of online content. But all claim to leverage AI, natural language processing (NLP) and machine learning (ML) technologies to evaluate whether a webpage's content is safe for advertising based on predefined safety categories (e.g., violence or alcohol) (Vargo, Hopp, and Agarwal 2024; Griffin 2023).

Correct classification of brand safety is critical for both advertisers and publishers, and yet, remains a huge challenge. On one hand, an article with safe content may be misclassified as unsafe. For example, an article discussing recent advances in breast cancer research can be classified as unsafe due to the presence of the keyword "breast". In such cases, advertisers lose opportunities and publishers face reduced revenues. On the other hand, an article with potentially unsafe content may be misclassified as safe, which results in advertisers wasting their budget and also risking their brand's reputation. One well-known example is YouTube's "Adpocalypse". Despite YouTube's keyword-based brand safety controls, major brands like AT&T, Pepsi, and Verizon pulled their ads from the platform after discovering their ads were placed alongside extremist or inappropriate content (Reed 2017). This highlights how misclassification can lead to financial and reputational damage for advertisers and publishers.

While professional editorial review reduces the likelihood of toxic language compared to user-generated platforms, it does not eliminate risk. News outlets regularly cover violence, pandemics, political extremism, and other sensitive topics that can trigger brand safety systems. In these cases, the underlying content is not 'unsafe' in the sense of misinformation or hate speech, but it is nevertheless subject to unsafe ratings due to its subject matter. Thus, news content presents a unique challenge: the coexistence of editorial standards with recurring coverage of high-risk topics. Misclassifying safe digital news content as unsafe will harm the struggling news publishers. Despite generating $12 billion in digital advertising revenue in 2019, an industry report estimated that news publishers across the United States failed to generate an additional $2.8 billion because of erroneous, or overly cautious, keyword-blocking practices that prevented safe content from being monetized (CHEQ 2020).

News publishers are increasingly reliant on digital advertising as print revenues decline (CHEQ 2020). However, even digital income is under strain: AI-driven search features have reduced referral traffic, with one 2025 report showing a 7% drop in visits from Google year over year (Kint 2025). This makes accurate brand safety classification vital, since

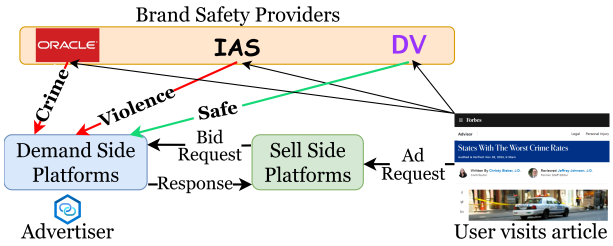

Figure 1: Three brand safety providers each classify an article at Forbes.com

mislabeling legitimate content as unsafe directly undermines publishers' ability to monetize their limited traffic.

It is noteworthy that the aforementioned examples were the result of keyword-based brand safety tools. At the time, there was hope that newer AI and contextual-based brand safety tools would do more accurate classifications (CHEQ 2020). Today, AI-based brand safety classification is prevalent. Our paper investigates whether this technological advancement is satisfactory.

The objective of this study is to analyze the effectiveness and consistency of the major brand safety providers: DV, IAS, and Oracle, to determine if they adequately protect the reputation of the advertisers' brand without losing advertising opportunities, and at the same time, prevent publishers' safe content from becoming demonetized.

Our study shows that different providers may give different classifications for the same article. Figure 1 shows how DV, IAS and Oracle categorize an article on Forbes' website. The article discusses crime, a topic that can be deemed unsafe. While Oracle classifies the content as unsafe and related to crime, IAS deems it unsafe with a low risk of violence, and DV classifies the article as safe. This illustrates the significant variation in the way brand safety tools interpret and classify the same content. If safety classifications are inconsistent, publishers would be confused about how to avoid demonetization, and advertisers do not know what filters to set to avoid placement alongside undesirable content.

Few studies have evaluated the effectiveness of brand safety tools. One evaluated brand safety classifications on Reddit (Vargo, Hopp, and Agarwal 2024), and another evaluated the brand safety classification of social media influencers' content (Bishop 2021). Our research differs from these works in several aspects. First, instead of studying user-generated content on social media platforms, we study the brand safety classification of news articles on reputable websites. News content was selected because it is a major concern for advertisers, and digital advertising revenue is a key profit source for publishers. Since news content encompasses a broad range of topics and styles, our findings are generalizable. Second, we evaluate multiple third-party brand safety providers, which are widely used in the online advertising industry. Furthermore, we study many dimensions of brand safety (e.g., violence, crime, alcohol) instead of just considering general toxicity, as in prior literature.

Inspired by the Forbes Ad Inspector[1], we built a data collection tool and collected brand safety data from 4,352 articles from popular news websites (e.g., nytimes.com, cnn.com, economist.com).

This study presents several major findings from our analysis of these 4,352 articles. First, brand safety is a prevalent issue for the news media that the public may not be aware of. Our research indicates that brand safety providers have a significant impact on the advertising ecosystem and the news media publishers. Of the articles with brand safety data, 43.99% were classified as "unsafe" by at least one of the three brand safety providers (Section 4.1). This impact is felt most by publishers that focus on breaking/general news, as we find that articles from these domains are more likely to be classified as unsafe (52.14% of articles).

Second, we identified significant inconsistencies in the overall article classifications (i.e., whether the content is unsafe for any reason) among the brand safety providers. Disagreement in an article's brand safety rating occurs in 22.98% of the cases where two providers rate the same article, and in 41.57% of cases where all three providers give ratings. Furthermore, we evaluate interrater reliability with Krippendorff's alpha, which adjusts for agreement that may occur by chance. The interrater reliability between all three, and each pair of, providers, falls well below conventional thresholds, indicating unreliable ratings (Section 4.1).

Third, an article may be unsafe for different categories, such as violence, adult content, hate speech, or drugs. Providers disagree not only on the assessment of an article's overall brand safety, but also on brand safety categories (see Table 1 for all categories). Different providers adopt different safety categories and different risk levels. Even when two providers consider the same safety category, their assessment of an article may be different. Disagreements between two and three providers occurred in 62.40% and 70.0% of cases, respectively, when at least one provider rated the article as unsafe for the category. Furthermore, the discrepancies vary significantly between different safety categories. The hate speech safety category, for example, was inconsistent 87% of the time; while the death/injury category was inconsistent 57% of the time (Section 4.2).

Furthermore, our study shows that the current brand safety classification is not reliable even within the ratings assigned by the same provider. For instance, IAS rates one article as unsafe regarding "Illegal and Recreational Drugs", but safe regarding "Drugs" (Section 4.2).

The absence of a common standard exacerbates may contribute to inconsistencies in brand safety classifications between providers. The Global Alliance for Responsible Media (GARM) previously provided a framework that defined a set of brand safety categories and risk levels that should be considered when ensuring brand safety to inform decision-making across platforms, advertisers, publishers, and brand safety providers. However, GARM was discontinued on August 9th, 2024 after a lawsuit from X (Gollasch 2024).

The Media Ratings Council (MRC) performs audits on several brand safety tools based on guidelines co-developed

[1]https://forbes.github.io/ad-inspector/

with the Interactive Advertising Bureau (IAB). However, these guidelines are not a comprehensive framework for brand safety and do not include specific definitions or examples of what constitutes safe or unsafe content (MRC 2018). The need for a unified and transparent framework is underscored by the news in October 2024 that the US government is interested in investigating brand safety scandals (Barwick 2024), signaling broader concerns about systemic issues in the industry.

The contributions of our study are threefold. First, to the best of our knowledge, this is the first work that assesses widely used brand safety tools in news articles. Second, our comprehensive analysis shows the prevalence of brand safety usage in the industry and, at the same time, the inconsistency of the brand safety classification system for news articles. Furthermore, our findings call for action on standardizing guidelines in the space of brand safety. Specifically, a unified framework of brand safety categories between providers is needed. Reliable brand safety tools that are not self-contradicting are required, and greater consistency among different brand safety tools should be expected. A more standardized and transparent system, with explainable classifications, would reduce risks for advertisers, ensure fairness for publishers, and promote trust within the digital advertising ecosystem.

2 Related Work

Brand safety has become a central concern for advertisers, as ensuring that ads do not appear next to harmful or inappropriate content is crucial to protecting a brand's reputation (Johnson, Voorhees, and Khodakarami 2023; Lee, Kim, and Lim 2021; Marvin and Meisel 2018; Griffin 2023; Hemmings 2021). Early approaches to brand safety relied largely on keyword-based filtering systems, which flagged content based on the presence of specific terms without accounting for context. Later, advanced methods that incorporate natural language processing (NLP) and machine learning (ML) were introduced to more effectively analyze the content (Vargo, Hopp, and Agarwal 2024; Griffin 2023).

Two works are related to ours in that they also evaluate brand safety ratings, but in different contexts. The most relevant one examines Reddit's brand safety classifications and compares them with toxicity scores calculated by the Perspective API. The authors find evidence of misclassifications, particularly on high-traffic forums, where potential ad revenues may influence safety classifications (Vargo, Hopp, and Agarwal 2024). The second most relevant work studies influencer management tools, which help pair advertisers with social media influencers and rate the brand safety of social media influencers' content. This study found that one such tool used a list of keywords to calculate a "family friendly rating" and was systematically biased against LGBTQ creators due to words like "queer" and "gay" falling on the keyword list (Bishop 2021). A third study indirectly evaluated brand safety tools by finding they were ineffective in fully protecting advertisers from misinformation websites that obscured their identity through "dark pooling" (Vekaria, Nithyanand, and Shafiq 2024).

Our study differs in several important aspects. We focus on ratings of news articles from multiple reputable websites (instead of user-generated content or misinformation websites) and collect ratings from multiple brand safety providers (instead of a single provider). Unlike previous work, we examine classification variation across numerous brand safety dimensions (e.g., crime, drugs, sexual content).

In the larger literature on brand safety, several articles try to address how placement alongside unsafe content can affect advertising effectiveness or brand reputation. One paper finds that it does not decrease effectiveness in the case of advertisements shown before "unsafe" Youtube videos (Bellman et al. 2018). Others have found that the advertisement's effectiveness and brand's reputation are negatively affected when advertisements are placed next to offensive or unsafe content (Johnson, Voorhees, and Khodakarami 2023; Lee, Kim, and Lim 2021). These studies are orthogonal to ours, but show the importance of brand safety for advertisers.

Several works are devoted to the development of image-based classifiers to classify images and videos as unsafe with respect to adult content (Vo, Cao, and Ton-That 2020; Wang et al. 2018; Jin, Wang, and Tan 2019; Ou et al. 2017), and additional sensitive safety categories such as gambling, guns, and gore (Vo, Cao, and Ton-That 2020). These works are also orthogonal to ours in that they attempt to build new classification systems for image and video content, while we audit existing systems that focus primarily on text content.

Our work relates to the broader literature on the shortcomings of algorithmic content moderation. Content moderation differs from brand safety classification in that it is applied to user-generated content instead of reputable news articles, and the goal is to protect users from toxic content (e.g., profanity or slurs) instead of advertisers from negative or inappropriate topics (e.g., violence). However, there are some similarities as the ultimate goal is to classify content based on its safety. One study finds that automated content moderation tools struggle to replicate nuanced human moderation decisions in Reddit (Samory 2021). Another study finds that LLMs can improve content moderation performance over traditional NLP tools (e.g., Perspective API), but that performance varies significantly between different LLMs (Kumar, AbuHashem, and Durumeric 2024). This finding echoes the inconsistencies we observe across brand safety providers. Prior research has also found that misclassifications resulting from algorithmic content moderation are unequally distributed across demographic groups. For example, the Perspective API tool was shown to unfairly penalize the content of LGBTQ influencers (Oliva, Antonialli, and Gomes 2021). Similarly, unfair decisions against the LGBTQ, black, and blind communities have been found on platforms such as YouTube, TikTok, and Twitter (Kingsley et al. 2022; Ball-Burack et al. 2021; Lyu et al. 2024; Harris et al. 2023).

Related literature has increasingly advocated for standardized, transparent, and contestable content moderation frameworks. Many researchers recognize that content moderation policy changes often happen silently, and advocate for greater transparency and clearer definitions of what is and is not allowed (Pisharody et al. 2025; Harris et al. 2023; Gomez et al. 2024). Several papers further propose explain-

able algorithms for content moderation so users get detailed answers for why their content was targeted (Kingsley et al. 2022; Gomez et al. 2024). Vaccaro et al. (2021) propose mechanisms for making decisions contestable by affected users. Our paper supports this growing consensus and calls for standardization, transparency, explainability, and mechanisms that allow for appealing incorrect decisions in brand safety classification protocols.

3 Method

This section presents the methodology used to collect and analyze data for this study. First, Section 3.1 provides a brief overview of how brand safety tools work. Then, we describe our system, illustrated in Figure 2, in Section 3.2 (*data collection*) and Section 3.3 (*data processing*). Data collection involves two specialized web crawlers to collect brand safety data in popular online news domains. Data processing involves several data cleaning steps to extract provider-specific brand safety details, which are then used in the analysis presented in Section 4. The cleaned dataset is available.

3.1 How Brand Safety Tools Work

Brand safety providers collect page text and metadata (e.g., title) for each article, and use proprietary AI and natural language processing (NLP) models, as well as rule-based heuristics to assign risk levels across predefined safety categories. These classifications are stored on the provider's servers and are used to protect advertisers from placing their ads next to unsafe content.

There are two primary ways for advertisers to obtain ad inventory: through real-time bidding (RTB) auctions, or through direct deals with publishers (e.g., guaranteed deals). Publishers often manage their ad inventory using Google Ad Manager (GAM), which helps publishers decide whether to show an ad from an advertiser who has a guaranteed deal, or to show the winning ad from an RTB auction. In an RTB auction, safety classifications are shared with the demand side platform (DSP), who represents the advertiser. The DSP will not bid on behalf of an advertiser if the impression does not meet the advertiser's predefined brand-safety requirements. Because this exchange happens server-side, these classifications are not visible to external observers.

In cases where publishers integrate brand safety tools directly with GAM via Google Publisher Tags (GPT), the process occurs on the client side, allowing us to observe the classifications. In this setup, the publisher's code issues an HTTP request to the brand safety provider when an article is loaded. The provider either retrieves a pre-computed rating from its database or performs the classification in real time and returns the result. These results are then appended to the HTTP request sent to GAM via the GPT tag, where they can influence the ad decision logic. The GPT integration thus enables both publishers (and researchers) to access page-level brand safety scores during article rendering.

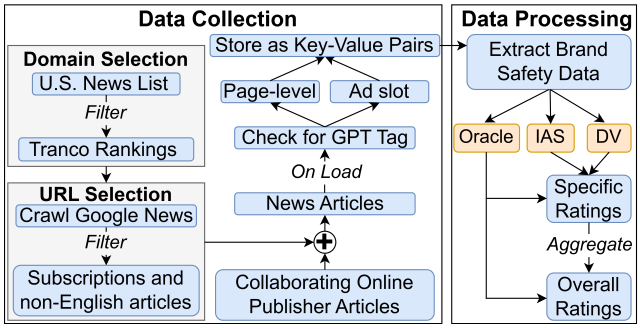

Figure 2: System Design

3.2 Data Collection Phase

The first phase of our system focuses on 1) collecting the URLs of articles from various news domains, and 2) collecting the GPT data for the articles.

Domain Selection We start with a list of 5,397 popular news domains developed by von Hohenberg et al. (2021), that was the union of Alexa's 1,000 top domains, domains frequently tweeted by US politicians, domains for local newspapers and television stations, and domains frequently visited by survey participants. This initial list does not rank the domains. We use the Tranco rankings (Le Pochat et al. 2019), which are widely adopted in academic research as an objective listing of web domain popularity, to rank the initial list of news domains by popularity.

We then visit the top 250 news domains. We only visit the top 250, so our analysis uncovers brand safety data patterns across the most influential online news sources. For each domain, we visit it using the Forbes Ad-Inspector tool, implemented as a bookmarklet. This tool displays all of the Google Publisher Tag (GPT) data related to the ads on the page. We use the tool to determine whether brand safety data is available on each website.

If the publisher chooses to use the brand safety providers tags' with its Google Ad Manager ad server (see Section 3.1), the brand safety data will be visible to us in the GPT data. Based on this filtering, we identified 55 top news domains that we could capture brand safety data from. In addition, our collaborating online publisher (who requested to be anonymous) provides data for four of their domains. Thus, we have a total of 59 target domains for analysis. Notably, websites choosing not to collect brand safety data does not affect whether advertisers can see this data. It only means that we can not access brand safety ratings on articles from these websites.

URL Selection Next, we collect webpages of news articles from the 59 target domains. The collaborating publisher provided a list of around 150 URLs for each of their four domains. For each of the remaining 55 target domains, we crawl 100 *news articles* (i.e., not other content such as home pages or author pages) using Google News. Specifically, we queried Google News with only the publisher's domain name (e.g., cnn.com) and selected the first 100 articles returned. Note that Google News links are click-through URLs, meaning they redirect to the actual article pages. The crawler follows these links and extracts the article URLs. For this purpose, we develop a JavaScript-based

web crawler using the Playwright automation framework[2] to crawl Google News and retrieve URLs for recent news articles on each target domain. To refine our dataset, we excluded redirects to subscription pages (6 domains that always redirect to a subscription page were removed) and non-English websites (2 domains were removed). Thus, 51 of the original 59 target domains remain, reflecting the filtering steps that were necessary to ensure the availability of brand safety data on news articles written in English. With roughly 100 URLs for each of the 47 domains we crawled and 150 URLs for each of the four domains from our collaborating publisher, we have a total of 5,241 article URLs (not including those that redirect to subscription pages) from 51 domains.

Collecting GPT data With the URL data set finalized, we proceed to the second phase, where we extract the *advertising data* embedded on each article's web page. For this purpose, we develop a second crawler that visits each article's URL and calls on the Forbes Ad Inspector to collect the GPT data related to the ads on the web page. Brand safety data can be stored at different levels: ad-level or page-level. We collect brand safety data from both levels because different providers and publishers choose to store the data differently. The extracted data are stored locally for further analysis. Due to technical constraints, such as metered paywalls, CAPTCHAs, broken links, and anti-crawling measures, we are unable to extract data from every article. As a result, we successfully retrieve GPT data from 4,352 of the 5,241 articles that we attempted to visit.

3.3 Data Processing

After data collection, we process the data collected from 4,352 articles to extract brand safety data whenever available. Brand safety classifications are successfully retrieved from 3,219 articles. This subset of 3,219 articles from 51 top news domains, constitutes our final dataset for analysis.

All brand safety data are stored in key-value pairs. We show examples of the collected (Figure 7) and processed (Figure 8) data in Appendix A. We collaborated with a Demand Side Platform (DSP) to get information on how brand safety data is stored and coded. This allows us to interpret brand safety data from the three brand safety providers: Oracle, IAS, and DV. Now we discuss the process of extracting brand safety data from each provider.

Oracle The extraction of Oracle brand safety data is the simplest. There is an overall binary rating for the article (that is, "safe" or "unsafe") stored under a key called *m_safety*. Oracle also has a list of flags indicating which specific safety categories (e.g., crime) the article is unsafe for. This list of flags is stored under a key called *m_categories*. The ten specific safety categories considered by Oracle are listed in the first row of Table 1 under the "Oracle" column. We extract the overall rating and the list of categories and store them in two separate variables. If the overall Oracle safety rating exists, we count the article as having Oracle brand safety data. In our data collection, there are 74 articles where the Oracle tag did not load properly and the value is "waiting". We do not count these articles in our analysis, leaving us with 1,053 articles that have Oracle brand safety data.

Integral Ad Science (IAS) There are several components for collecting IAS brand safety data. In total there are 29 safety categories. The 11 safety categories in the first row of the "IAS" column in Table 1 are binary flags only indicating the presence of the category. These 11 categories are stored under the *ias-kw* key. The other 18 safety categories can take on risk levels. The seven safety categories in the middle row of the "IAS" column in Table 1 are stored under separate keys. They can take on values of "veryLow", "low", "medium", or "'high". The "veryLow" value is the default value and means the article is safe with respect to the specific category, and the other values represent varying levels of risk with regard to the unsafeness of the article. We extract these seven safety categories and store each as their own variable. The remaining 11 safety categories (shown in the last row of the column "IAS" in Table 1) are also stored under the *ias-kw* key. These 11 safety categories have risk levels of low, medium, or high. However, they are different from the safety categories in the middle row of the "IAS" column in Table 1 as they have an additional classification that specifies whether the unsafe content is related to news, entertainment or video games.

All brand safety values under the *ias-kw* key are coded as a 7-digit number with '_PG' appended. We decode these numbers using our DSP access. Some values stored under *ias-kw* are just codes for keywords or contextual targeting categories that do not relate to brand safety, and thus, are removed. We store the list of decoded brand safety categories for each article in a variable called *ias-kw-decoded*.

IAS typically does not provide an overall safety rating for the article. Advertisers need to choose the safety categories they want to avoid (e.g. terrorism and politics), and check the safety rating with respect to those categories. For the purpose of comparison with Oracle and DV, we construct an overall binary safety rating for each article: if the *ias-kw-decoded* is empty and all seven of the separately stored category ratings take on a "veryLow" value, we consider the article to be rated safe by IAS, otherwise it is rated unsafe. For one domain (theglobeandmail.com), IAS instead stores values under the *ias_admants* key. For this publisher, IAS does not consider specific safety category ratings, but rates its overall safety (through the presence/absence of a "brand_unsafe" string).

As long as any of the IAS key-value pairs exist, we consider the article to have IAS brand safety data. This results in 1,666 articles with IAS brand safety data. For most articles, both the *ias-kw* and seven variables in row two of Table 1 are scored. However, 80 articles only have *ias-kw* data and 232 articles have the *ias-kw* unscored.

DoubleVerify (DV) The DV brand safety data is stored as a list of 8-digit codes under the *BSC* or *bsc* key. We use our DSP access to decode these 8-digit codes. We only retain the codes that are brand safety categories, and we store the list of decoded categories in the "bsc-decoded" variable. DV provides both specific category ratings and an overall article

[2]https://github.com/microsoft/playwright

Levels	Oracle	IAS	DV
Binary	Drugs, Death/Injury, Crime, Military, Adult, Arms, Terrorism, Hate Speech, Tobacco, Obscenity	Politics, Pop Culture, Animal Cruelty, Discrimination, Conspiracy Theories, Infectious Diseases, Protests, Misinformation, Pollution, Smoking, Natural Disasters	Any Moderate Content, Any Severe Content, Occult, Gambling, Pharmaceutical, Financial, Celebrity Gossip
Low / Medium / High		Adult, Alcohol, Illegal Downloads, Drugs, Hate Speech, Offensive Language, Violence	Violence, Death/Injury, Drug Abuse, Crime, Adult/Sexual, Human-made Disaster, Terrorism, Hate Speech/Cyberbullying, Alcohol, Natural Disasters, Vehicle Disasters, Profanity
(Low / Medium / High) + (News / Entertainment / Video Games)		Death/Injury/Military, Crime, Sensitive Social Issues, Terrorism, Incitement of Hatred, Sexual Content, Arms/Ammunition, Cyber Security, Illegal Drugs, Online Piracy, Obscenities	

Table 1: Brand Safety Categories by Provider

safety rating. If any brand safety category is present, there is an additional code for the "Any moderate content" category. There is also a code for "Any severe content", but we see this only twice in our dataset. If the "Any moderate content" code is present, we consider the article to be rated as unsafe. DV considers 17 safety categories. As shown in the first row of the "DV" column in Table 1, there are five categories that are only recorded for presence, and do not have an associated risk level. The second row lists the remaining 12 safety categories, where risk ratings are low, medium, or high. As long as the "BSC" or "bsc" key is not empty, we consider the article to have been rated by DV for brand safety. This results in 1,261 articles.

4 Results

Our data allows us to explore brand safety ratings from two perspectives: overall classifications and specific classifications by safety categories. Overall classifications categorize an article as "safe" or "unsafe" on a holistic level. Exploring brand safety at this level is important because some advertisers choose not to advertise on a web page where content is rated "unsafe" according to any category. However, it is also important to analyze brand safety ratings for specific safety categories (e.g., crime or hate speech) because different brand safety providers consider different categories to define overall brand safety, and thus their overall brand safety ratings may not be directly comparable. Next, we present the results and insights for the overall and specific category classifications of the 3,219 articles obtained after processing the data as described in Section 3.3.

4.1 Overall Classifications

The analysis of brand safety ratings for the overall classification relies on the definitions of safe and unsafe that we derived for each provider, as discussed in Section 3.3. We find that 1,416 of the 3,219 articles in our sample (43.99%) had an unsafe rating from at least one of the brand safety providers. This indicates that brand safety is having a large impact on the digital advertising industry. Our findings show that articles on domains publishing general/breaking news stories are more likely to be classified as unsafe compared with domains on specific news topics (52.14% vs. 32.67%). We also found that some brand safety providers are more stringent than others and, most importantly, that classifications are highly inconsistent between providers.

Oracle The 1,053 articles with Oracle brand safety data are from 16 domains (see Section 3.3). Oracle rates 234 (22.22%) of these articles as unsafe. The breakdown of these

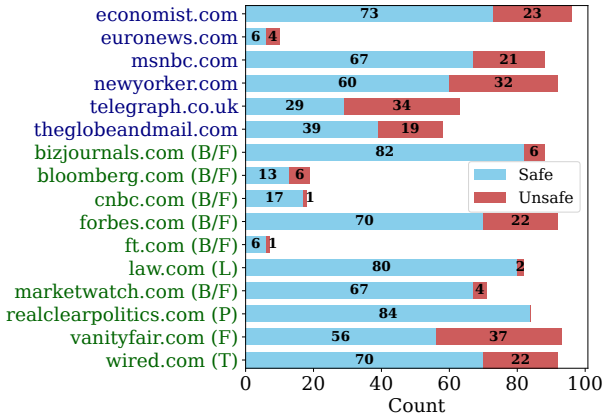

Figure 3: Oracle ratings by domain
Note: Navy=Breaking/General News, Green=Specific Topic, B/F=Business/Finance, L=Law, P=Politics, T=Technology, F=Fashion

safe and unsafe ratings by domain is shown in Figure 3. We categorize domains into different categories, such as general, breaking news, or a specific topic such as business, laws, politics, etc. The details of labeling a domain into news categories can be found in Table 3 in Appendix C. There is a big difference in the percentage of articles rated as "safe" for domains that focus on a special topic (e.g., business/finance or law) versus general/breaking news. Articles on domains that focus on a specific topic are rated as safe by Oracle 84.37% of the time. One domain (realclearpolitics.com) has all safe ratings, and another four domains, all focusing on a specific topic, have over 90% safe ratings. On the other hand, only 67.32% of articles on domains in the general/breaking news category are rated as safe. This may be due to the fact that breaking news publishers are more likely to publish articles about inherently unsafe content (e.g., natural disasters, terrorist attacks, or wars). Of the 16 domains with ratings from Oracle, there is only one domain (telegraph.co.uk, a general news website) where safe ratings are in the minority (46.03%).

Integral Ad Science (IAS) The 1,666 collected articles with IAS brand safety data are from 31 domains (see Section 3.3). IAS rates 895 (53.72%) of these articles as unsafe. The breakdown of these safe and unsafe ratings by domain is shown in Figure 4. It is very different from Oracle's ratings (see Figure 3) in that there are 16 domains with more articles rated unsafe than safe by IAS. There is one domain with only safe ratings (investors.com), but there is also a domain with only unsafe ratings (dailymail.co.uk). The trend that articles from general/breaking news publishers are more likely to be rated unsafe still holds true for IAS. Articles on specific news topics domains are rated safe 55.77% of the time, while articles on general/breaking news domains are rated safe only 38.27% of the time.

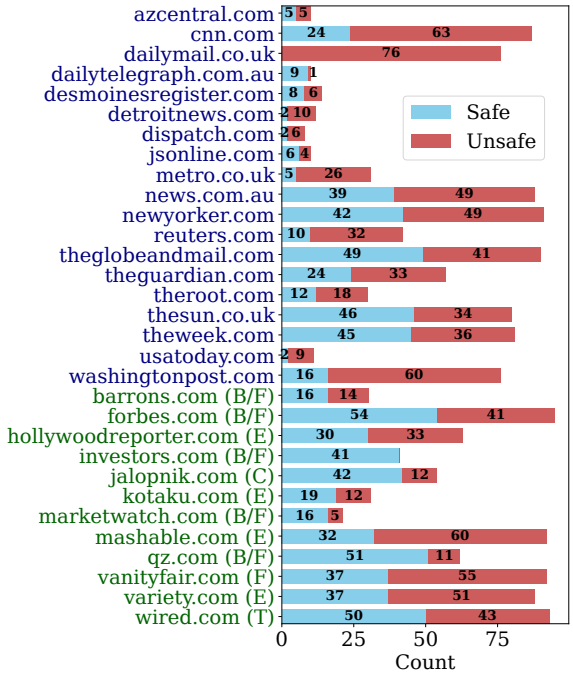

Figure 4: IAS ratings by domain
Note: Navy = Breaking/General News, Green = Specific Topic, B/F = Business/Finance, E = Entertainment, C = Cars, T = Technology, F = Fashion

DoubleVerify (DV) The 1,261 articles collected with DV brand safety data are from 16 domains (see Section 3.3). DV rated 448 of these articles (35.27%) as unsafe. There are two domains with over 90% of the articles rated as safe, both of which specialize in business/finance news. There are five domains where the majority of articles are rated unsafe, and four of them are general news domains; the other is vulture.com, which focuses on entertainment. The domains that focus on a specific topic have articles classified as safe 77.9% of the time, while the general/breaking news domains have articles classified as safe only 57.95% of the time.

Comparison between brand safety providers We find that in general IAS is most stringent, with the highest percentage of articles classified as unsafe, with DV second highest, and Oracle lowest. This may be due to IAS considering 29 brand safety categories, while DV and Oracle consider only 17 and 10, respectively (see Table 1).

Next, we analyze the 672 articles that have multiple providers' brand safety ratings and compare different

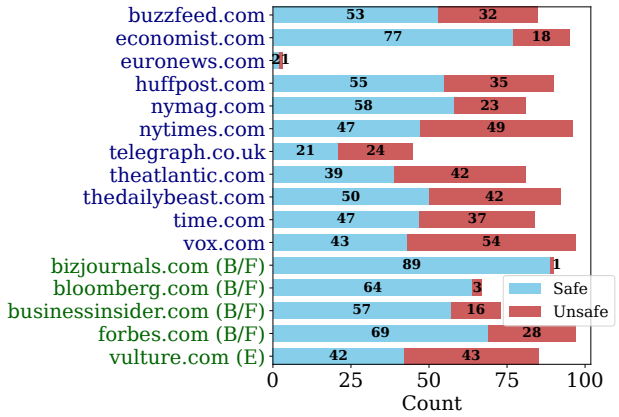

Figure 5: DV ratings by domain
Note: Navy = Breaking/General News, Green = Specific Topic, B/F = Business/Finance, E = Entertainment

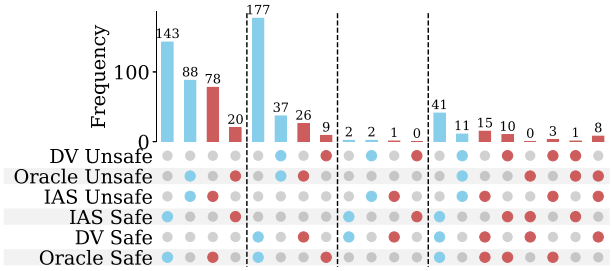

Figure 6: Brand Safety Disagreements

providers' ratings. There are 329 articles rated by just Oracle and IAS, 249 articles rated by just Oracle and DV, 5 articles rated by just IAS and DV, and 89 articles (all on forbes.com) rated by all three providers. In Figure 6, we use an UpSet style plot (Lex et al. 2014) to show how each of these 672 articles is rated for overall safety by the different providers. The vertical dotted lines separate which providers are being compared in each section (e.g., the first section compares Oracle and IAS). Blue bars indicate agreement between providers, while red bars indicate disagreement. In total there is disagreement in 22.98% of the 583 articles with ratings from only two providers. There is even higher disagreement when all three providers generate ratings: 41.57% of the 89 articles.

The first four bars in Figure 6 break down (dis)agreements between Oracle and IAS on 329 articles rated by just these two providers. Disagreement is high (29.79%). Such disagreements are not just due to different levels of tolerance for "unsafe" content. Across the entire dataset, IAS rates articles as unsafe at the highest rate, and Oracle at the lowest rate. However, there are 21 articles rated safe by IAS and unsafe by Oracle (including 1 article rated unsafe by all three providers). Furthermore, we calculated Krippendorff's alpha (α) to demonstrate the poor agreement between providers. Calculating interrater reliability is a more accurate measure of inconsistency than raw percentages because it accounts for agreement by chance. This is particularly important in our dataset where safe classifications are more likely than unsafe classifications. IAS and Oracle had an interrater reliability of only 0.42, far below the recommended threshold of 0.8, and even below the absolute minimum acceptable threshold of 0.67 for ratings that are only used for tentative conclusions (Krippendorff 2004).

The second set of four bars in Figure 6 break down the (dis)agreements on 249 articles rated by just Oracle and DV. Disagreement is lower (14.06%), but this is partially due to both providers rating the majority of articles as safe. This indicates potential agreement by chance. The interrater reliability is higher than any other pairing of brand safety providers ($\alpha = 0.53$), but is still far below the recommended and minimum thresholds of 0.8 and 0.67, respectively.

The third section of bars in Figure 6 break down the (dis)agreements on the five articles rated by just IAS and DV. There are so few articles because IAS and DV are only both present on 94 forbes.com articles in our sample and MOAT also rates 89 of these articles (shown in the final section of bars in Figure 6). Across all 94 articles rated by IAS and DV, the disagreement is high (37.23%) and the interrater reliability is very low ($\alpha = 0.19$).

Classifications of the 89 articles rated by all three providers is shown in the final section of Figure 6. Disagreement between the three providers is high (41.57%), and the interrater reliability is low ($\alpha = 0.35$).

4.2 Specific Classifications

The results from Section 4.1 show the inconsistency of brand safety tools in assessing the overall safety of an article. However, this inconsistency can be partially explained by variation in the number and nature of safety categories assessed by each provider. Additionally, some advertisers may only block a few specific brand safety categories, and do not care about the overall safety rating. For example, a beer company may care specifically about not advertising on news articles about alcoholism, alcohol abuse, drug abuse, or drunk driving accidents, and set up brand safety filters accordingly. Thus, in this section, we examine the inconsistency in brand safety classifications with respect to a particular safety category. A descriptive analysis of which safety categories are most frequently occurring under each provider in our dataset is presented in Appendix B.

Comparison between brand safety providers Each provider considers different safety categories and different risk levels (see Table 1). To compare the providers, we identify several brand safety categories that exist in the list of at least two brand safety providers (regardless of risk level). We make these pairings conservatively, and only pair categories from different providers when the category is not ambiguous. For example, we do *not* pair the "Drugs" category for IAS with the "Drug Abuse" category from DV because drug abuse content is more severe than drug-related content. We do, however, allow for combining two categories under one brand safety provider to make a match. For example, IAS has a "Death, Injury or Military Conflict" category. While Oracle does not have this exact category, it does have "Military" and "Death & Injury" categories. We combine these two categories for Oracle to allow comparison with IAS. The full list of pairings is displayed in Table 2. For each safety category listed, we can tell which providers are compared from the "Oracle", "IAS", and "DV" columns. The "T" values indicate which provider is being compared for the specific category.

Category	Oracle	IAS	DV	Inconsistencies	α
Arms/Ammunitions	T	T	F	6/9 (66.67%)	0.49
Drugs	T	T	F	8/9 (88.9%)	0.19
Adult	T	T	F	14/23 (60.87%)	0.54
Adult/Sexual	F	T	T	6/9 (66.7%)	0.47
Natural Disasters	F	T	T	1/1 (100%)	0.0
Terrorism	T	T	T	2/3 (66.67%)	0.66
Terrorism	T	T	F	12/18 (66.67%)	0.48
Terrorism	T	F	T	5/12 (41.67%)	0.73
Terrorism	F	T	T	2/3 (66.67%)	0.49
Crime	T	T	T	5/7 (71.43%)	0.56
Crime	T	T	F	37/61 (60.66%)	0.50
Crime	T	F	T	22/41 (53.66%)	0.6
Crime	F	T	T	3/7 (42.86%)	0.7
Death/Injury	T	F	T	24/42 (57.14%)	0.56
Death/Injury/Military	T	T	F	35/63 (55.56%)	0.55
Hate speech	T	T	F	27/31 (87.10%)	0.19
Violence	F	T	T	19/26 (73.08%)	0.31
Alcohol	F	T	T	1/2 (50%)	0.66
Obscenities	T	T	F	2/2 (100%)	0.00
Total	-	-	-	**231/369 (62.60%)**	-
-	**T**	**T**	**F**	**141/216 (65.28%)**	**0.37**
-	**T**	**F**	**T**	**51/95 (53.68%)**	**0.63**
-	**F**	**T**	**T**	**32/48 (66.67%)**	**0.44**
-	**T**	**T**	**T**	**7/10 (70%)**	**0.61**

Table 2: Comparing Brand Safety Ratings for Specific Categories

After identifying the set of safety categories supported by all brand safety providers, we compare their respective classification outcomes for each shared category. For each category, we locate all articles with ratings from both providers and compute interrater reliability using Krippendorff's alpha (see "α" in Table 2). For interpretability, we also compute the inconsistency as a percentage. The majority of articles are not related to any given category and thus providers are in agreement by default. Therefore, we only consider articles with at least one unsafe rating when reporting disagreement as a percentage.

Since there is no risk level provided for many safety categories, we consider two providers to agree when they both rate the article as unsafe for the target category, irrespective of risk level (e.g., if one provider rates unsafe with low risk and the other rates unsafe with high risk, we consider them in agreement). Thus, we are making a conservative estimate of the inconsistency between brand safety tools. While this analysis does not always enable us to reveal the provider that misclassified the article, manual evaluation of some articles shows that all providers are inaccurate in some cases.

In total, there are 231 inconsistencies on 369 instances (62.60%). There is an inconsistency on 7 of the 10 instances with a rating from all three providers. The average interrater reliability across all comparable categories for the three providers (α=0.53) and each pair of providers is always below the minimum threshold of 0.67. The highest average agreement (α=0.63) is between Oracle and DV, which can be compared for three categories. The lowest average interrater reliability (α=0.37) is between IAS and Oracle, which can be compared across eight categories.

Digging in deeper, Table 2 shows 141 disagreements between IAS and Oracle, 65.28% of all 216 instances where IAS and Oracle can be compared. There were 75 times where only Oracle flagged the category as unsafe, and 66 times only IAS flagged a category unsafe, with 39 times being low risk, 26 times being medium risk, and once being high risk. In the most egregious example, where IAS rated an article as high risk for drug content while Oracle rated it as safe, it appears IAS is at fault because the article covers an update to the Samsung Galaxy phones. Our manual inspection of this article revealed no reason for this. As an example where Oracle is at fault, an article in *The New Yorker* about "Black Identity" is labeled by Oracle as being unsafe and related to crime.

Table 2 shows 51 instances of an inconsistency between Oracle and DV ratings, 53.68% of the 95 instances that Oracle and DV can be compared. Most disagreements are due to Oracle being more stringent, flagging a category as unsafe when DV rates it as safe (41 times). As an example, an article about the Hezbollah pager explosions cites 9 dead and 2,750 injured, yet DV does not rate the article unsafe for "Death & Injury".

Finally, there are 32 cases of disagreement between IAS and DV, 66.67% of the 48 instances that can be compared. IAS is a more stringent provider, flagging 26 out of the 32 instances as unsafe with a risk level of low 13 different times, medium 11 times, and high twice. For the other 6 disagreements (where DV rates the article as unsafe and IAS does not), DV rates the risk level as low once and medium five times.

We can also look at the two safety categories (Terrorism and Crime) in Table 2 where all three brand safety providers can be compared. For "Terrorism", there are only three articles for comparison and two are disagreements. For "Crime", there are seven articles for comparison, and five have a disagreement. In three of the disagreements only Oracle rates the content as unsafe.

Comparison among safety categories We further analyze the inconsistencies across different safety categories, as shown in Table 2. The interrater reliability for all specific categories falls below the recommended reliability threshold of 0.8, but there are two instances (IAS and DV rating crime, and Oracle and DV rating terrorism) where the interrater reliability exceeds the minimum acceptable threshold of 0.67. The remaining 17 categories have interrater reliability scores that signal unreliable ratings, which may indicate that the corresponding safety category is harder to classify. Under this assumption, our results illustrate which safety categories need more concrete definitions in an updated framework of brand safety. Brand safety providers should also prioritize improving classification accuracy within the safety categories with a higher level of inconsistency.

Comparisons within brand safety providers We also examine the consistency of behavior within a brand safety provider. For IAS, sometimes the category ratings are in conflict with each other. For example, there are separate "Adult" and "Sexual Content"safety categories, and we expect these two categories to have similar ratings. There are 179 articles where both categories are considered and at least one of them has a low, medium, or high risk rating. Of those 179 articles, there are:

- 12 articles where both categories are rated unsafe with the same risk level.
- 12 articles where both categories are rated as unsafe, but with different risk levels.
- 155 articles where one category is safe and one is unsafe.

We consider some of these inconsistencies to be extreme. For example there are 36 articles where one category is rated as safe and the other is unsafe with a risk level higher than low (35 with medium risk and 1 with high risk). Looking into these 36 articles, we find that 12 of them discuss sexual assault, sex trafficking, or rape, and yet have a safe rating for either "Adult" or "Sexual content". In contrast, there is also an article about the construction plans for a highway in Arizona classified as medium risk for adult content.

There is a similar trend occurring for the "Drugs" and "Illegal and Recreational Drugs" categories. There are 46 articles where both safety categories are considered and at least one of the two drug-related categories has a low, medium, or high risk rating. There are only two articles where both categories have an unsafe rating, and the risk levels for the two categories are not the same in either case. For the other 44 articles, there are 33 times where "Illegal and Recreational Drugs" is rated as safe and "Drugs" is not (30 low risk, 2 medium risk, 1 high risk). There are 11 articles

where "Drugs" is rated as safe and "Illegal and Recreational Drugs" is not (all medium risk). Looking at the 14 articles with medium or high risk we again find that some of these articles are not ambiguous. Three of the articles mention hard drugs such as ketamine or ecstasy, and one of the articles is the aforementioned article about Samsung Galaxy phones that is rated high risk for drug content.

For DV, the aggregate category for severe content is present each time that there is high risk for alcohol or profanity (one of their occurrences is on the same article). However, DV does not classify the seven articles with high risk for "Adult & Sexual" content as "severe content". It is unclear why being rated as unsafe with high risk for some categories and not for others would classify an article as "severe content".

5 Discussion

Each provider approaches brand safety with a different proprietary algorithm and set of brand safety categories. Combined with a lack of standardization and regulation, inconsistencies occur as expected. These problems are not limited to breaking news media. Our data show that inconsistency extends across many different categories such as entertainment, health, and finance. Inconsistent ratings are especially harmful to publishers as they can not reliably figure out how to prevent unsafe classifications, nor the resulting decline in advertising revenue. Advertisers can mitigate some of the harms of inconsistent ratings by relying on a single provider (publishers do not have this choice); however, we also find inconsistency within the ratings of a single provider.

We also manually identified instances of clearly incorrect brand safety ratings. These misclassifications carry serious consequences: advertisers risk damaging their brand when their ads are served next to harmful content misclassified as safe, and publishers lose revenue when valuable, contextually appropriate content is mislabeled as unsafe. The consequences of misclassification in content moderation are complex, and prior research has shown that such misclassifications typically harm marginalized groups disproportionately (Kingsley et al. 2022; Oliva, Antonialli, and Gomes 2021; Gomez et al. 2024; Harris et al. 2023). This makes the fact that an article about "Black Identity" was labeled as unsafe and related to crime in our dataset unsettling. Although we did not have enough relevant articles to draw concrete conclusions, future research should investigate this potential problem more thoroughly.

Recommendations. While our findings point to serious flaws in the current digital advertising landscape, advertisers still value safely-aligned ads placements, and publishers need monetization tools that protect brand integrity without sacrificing legitimate content.

Our findings call for a standardized, transparent framework for brand safety classification, which is essential for ensuring consistency, accuracy, and trust for a sustainable digital advertising ecosystem on the open web. This framework should include:

- Clear definitions of safety categories, with concrete examples of what constitutes risks at different levels
- Auditable classification processes, where providers agree to regular third-party evaluations in exchange for industry accreditation,
- An appeal mechanism for publishers and advertisers to contest misclassifications
- A public benchmark dataset, grounded in human-labeled brand safety ratings, to evaluate accuracy
- Explainable brand safety ratings

While brand safety providers may initially resist such measures, advertisers and publishers, who ultimately fund these tools, hold the leverage. If they collectively demand compliance with a standardized framework, non-compliant brand safety providers may be pushed out of the market. Importantly, such a system would still allow differentiation: providers could offer better classification quality and additional features such as specialized categories, fine-grained risk levels, or dynamic pricing models.

Our study advances academic and technical discourse on content classification and calls for deeper engagement from the Web, natural language processing, and machine learning communities to build culturally sensitive, explainable brand safety models, an urgent need amid increasing regulatory scrutiny (Barwick 2024).

Limitations. While our study presents valuable findings of inconsistencies in brand safety classification among leading providers, several methodological and conceptual limitations should be noted.

Although our analysis is limited to English-language news articles, the brand safety systems we evaluate are not confined to this domain. These tools are widely deployed across international markets and are applied to various formats, including video, podcasts, and user-generated content. Consequently, the methodological concerns we identify—classification inconsistency and the absence of ground-truth benchmarking—are likely to extend beyond our dataset and may affect brand safety assessments across languages, formats, and cultural contexts. Investigating how classification systems perform across these diverse types of media would provide a more comprehensive understanding of brand safety practices in the digital ecosystem.

Second, our primary analytical focus is on classification inconsistencies between brand safety providers rather than classification accuracy. This is due to the absence of a ground-truth dataset of brand safety labels against which we could benchmark provider's output. Without independent, human-verified assessments of content safety, we cannot definitively determine whether any given provider's classification is correct or incorrect. We call for future research to develop such a benchmark using human expert annotation. Developing a human-annotated benchmark to evaluate the accuracy of brand safety classifications remains an important avenue for future work, and we are actively exploring best practices for designing such a dataset in consultation with industry providers. Recent work indicates that carefully calibrated moderation interventions can steer users away from extremist movements (Russo et al. 2025), highlighting the broader social implications of reliable classification of safe content.

Third, our dataset only includes webpages for which their publishers choose to collect brand safety data from the providers. This introduces a potential selection bias. Consequently, our findings on the proportion of articles labeled unsafe may be influenced by publishers' self-selection, potentially skewing our understanding of the overall coverage of brand safety. Our analysis assumes that the publisher's decision to collect brand safety data is independent of the safety classification provided by a brand safety provider, a conceptual constraint that we acknowledge and plan to test in future research.

6 Conclusions and Future Work

This paper presented an empirical analysis of brand safety classifications from three industry leading providers, DoubleVerify (DV), Integral Ad Science (IAS), and Oracle. Our findings show that brand safety has a large effect on the digital advertising industry as there was an unsafe rating from at least one provider on 43.99% of the articles in our dataset. The frequency of unsafe ratings is not a problem in itself, but the lack of standardization and consistency in the ratings is. We found inconsistent classifications between two providers in both overall safety assessments (22.98%) and individual risk categories (62.40% of the time where at least one unsafe rating is present). Low interrater reliabilities indicate that both the overall safety, as well as the specific category-level, ratings from brand safety providers are unreliable. These inconsistencies are not isolated anomalies, but indicate deeper issues stemming from inherent technical challenges, proprietary algorithms and varying interpretations of what constitutes "unsafe" content. Finally, we showed that IAS sometimes produces internally inconsistent ratings. For instance, there are 46 articles where IAS rated as safe for the 'Drugs' category but as unsafe for the 'Illegal and Recreational Drugs' category, or vice versa.

Such inconsistencies carry serious consequences and undermine the trust in brand safety technologies that are central to the operation of the digital advertising ecosystem. If publishers and advertisers cannot correctly identify unsafe content, publishers cannot mitigate the revenue losses associated with content that is classified as unsafe, and advertisers may reduce their digital advertising budget to protect their brand's reputation. Furthermore, some inconsistencies are due to blatant misclassifications, which directly harm publishers' revenues when safe content is incorrectly flagged as unsafe, and advertisers' reputations when their ads are shown next to unsafe content that is misclassified.

Although our dataset is primarily made up of English-language web articles, the issues that we have identified of inconsistent classification of safe content both within and between providers are fundamental to inherent technical challenges in classification and lack of standardization. Therefore, we believe that our results are broadly generalizable across different content, geographies and media types.

To address these issues, we advocate for the implementation of a standardized, transparent, explainable, and auditable brand safety framework. Such a framework would include clearly defined safety categories, routine third-party audits, explainable classifications, and an appeal process for misclassified content. By adopting a shared protocol, the industry can move toward more reliable and fair evaluations that serve the interests of advertisers and publishers.

Our findings also raise important areas for future research. Content safety classification poses inherent technical challenges to natural language understanding. The necessity of evaluating words in their context (e.g., "shoot" has different meanings in a sports article versus a breaking news story about a violent criminal) lends itself well to large language models. However, open research questions remain about how to effectively handle hierarchical classification (e.g., assigning low, medium, or high risk levels to certain categories), mitigate potential biases against specific dialects or writing styles, and generate explanations for each classification that are both faithful and plausible.

Acknowledgements

This work was supported by the NSF under Grants No. CNS 2237328 and DGE 2043104, as well as by the Martin Tuchman'62 Chair Endowment and the Leir Foundation. ChatGPT and Writefull were utilized to refine some sentences for clarity, grammar, spelling, and punctuation.

References

Ball-Burack, A.; Lee, M. S. A.; Cobbe, J.; and Singh, J. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Barwick, R. 2024. DOJ, NCIS ask ad executives about brand-safety companies. *Marketing Brew*. https://www.marketingbrew.com/stories/2024/10/11/doj-ncis-brand-safety-google-integral-ad-science-doubleverify.

Bellman, S.; Abdelmoety, Z. H. S.; Murphy, J.; Arismendez, S.; and Varan, D. 2018. Brand Safety: The Effects of Controversial Video Content on Pre-Roll Advertising. *Heliyon*, 4(12).

Bishop, S. 2021. Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias. *Social Media + Society*, 7(1): 1–13.

Business Insider. 2024. DoubleVerify and Integral Ad Science face fresh scrutiny from brand safety watchdogs. https://www.businessinsider.com/doubleverify-integral-ad-science-face-fresh-brand-safety-scrutiny-2024-6.

CHEQ. 2020. The Economic Cost of Keyword Blacklists for Online Publishers. Technical report, University of Baltimore.

Elmore, J. 2024. Who are DoubleVerify Competitors: A Closer Look at the Top Contenders. *TheTechyLife*. https://thetechylife.com/who-are-doubleverify-competitors/.

Gollasch, G. 2024. WFA makes 'difficult decision' to discontinue GARM following X lawsuit. *Marketing Week*. https://www.marketingweek.com/wfa-suspend-garm-x-lawsuit/.

Gomez, J. F.; Machado, C.; Paes, L. M.; and Calmon, F. 2024. Algorithmic Arbitrariness in Content Moderation. In

The 2024 ACM Conference on Fairness, Accountability, and Transparency.

Griffin, R. 2023. From Brand Safety to Suitability: Advertisers in Platform Governance. *Internet Policy Review*, 12(3).

Harris, C.; Johnson, A. G.; Palmer, S.; Yang, D.; and Bruckman, A. 2023. 'Honestly, I Think TikTok Has a Vendetta Against Black Creators': Understanding Black Content Creator Experiences on TikTok. In *Proceedings of the ACM Human-Computer Interaction*.

Hemmings, M. 2021. Ethical Online Advertising: Choosing the Right Tools for Online Brand Safety. *Journal of Brand Strategy*, 10(2): 109–120.

Jin, X.; Wang, Y.; and Tan, X. 2019. Pornographic Image Recognition via Weighted Multiple Instance Learning. *IEEE Transactions on Cybernetics*, 49(12): 4412–4420.

Johnson, R. W.; Voorhees, C.; and Khodakarami, F. 2023. Is Your Brand Protected? Assessing Brand Safety Risks in Digital Campaigns. *Journal of Advertising Research*, 63(3): 205–221.

Kingsley, S.; Sinha, P.; Wang, C.; Eslami, M.; and Hong, J. I. 2022. "Give Everybody [..] a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proceedings of the ACM on Human-Computer Interaction*, 6: 1–37.

Kint, J. 2025. Facts: Google's push to AI hurts publisher traffic. https://digitalcontentnext.org/blog/2025/08/14/facts-googles-push-to-ai-hurts-publisher-traffic/.

Krippendorff, K. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3): 411–433.

Kumar, D.; AbuHashem, Y.; and Durumeric, Z. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. *Proceedings of the 18th International AAAI Conference on Web and Social Media*.

Le Pochat, V.; Van Goethem, T.; Tajalizadehkhoob, M., Samaneh Korczyński; and Joosen, W. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019.

Lee, C.; Kim, J.; and Lim, J. S. 2021. Spillover Effects of Brand Safety Violations in Social Media. *Journal of Current Issues & Research in Advertising*, 42(4): 354–371.

Lex, A.; Gehlenborg, N.; Strobelt, H.; Vuillemot, R.; and Pfister, H. 2014. UpSet: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, 20(12): 1983–1992.

Lyu, Y.; Cai, J.; Callis, A.; Cotter, K.; and Carroll, J. M. 2024. 'I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.': A Study of Blind TikTokers' Content Moderation Experiences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Marvin, G.; and Meisel, S. 2018. Protecting the brand in the era of fake news: Why brands need advertisement verification tools. *Journal of Digital & Social Media Marketing*, 5(4): 322–331.

MRC. 2018. MRC Supplement to IAB Guidelines for the Conduct of Ad Verification: Enhanced Content Level Context and Brand Safety. https://www.mediaratingcouncil.org/sites/default/files/Standards/MRC%20Ad%20Verification%20Supplement-%20Enhanced%20Content%20Level%20Context%20and%20Brand%20Safety%20%28Final%29.pdf.

Oliva, T. D.; Antonialli, D. M.; and Gomes, A. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*.

Ou, X.; Ling, H.; Yu, H.; Li, P.; Zou, F.; and Liu, S. 2017. Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-Coarse Strategy. *ACM Transactions on Intelligent Systems and Technology*, 8(5): 1–25.

Pisharody, N.; Norton, S.; Greene, K.; and Shapiro, J. 2025. Changes in YouTube's Content Moderation Policy Had Little Detectable Impact on Election Denial Content. *Proceedings of the 19th International AAAI Conference on Web and Social Media*.

Reed, B. 2017. Google's bad week: YouTube loses millions as advertising row reaches US. *The Guardian*. https://www.theguardian.com/technology/2017/mar/25/google-youtube-advertising-extremist-content-att-verizon.

Russo, G.; Styczen, M.; Horta Ribeiro, M.; and West, R. 2025. Does Content Moderation Lead Users Away from Fringe Movements? Evidence from a Recovery Community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 1719–1734.

Samory, M. 2021. On Positive Moderation Decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 585–596.

Vaccaro, K.; Xiao, Z.; Hamilton, K.; and Karahalios, K. 2021. Contestability for Content Moderation. In *Proceedings of the ACM on Human-Computer Interaction*.

Vargo, C. J.; Hopp, T.; and Agarwal, P. 2024. Balancing Brand Safety and User Engagement in a Two-Sided Market: An Analysis of Content Monetization on Reddit. *Journal of Current Issues & Research in Advertising, 45(2), 242–256*.

Vekaria, Y.; Nithyanand, R.; and Shafiq, Z. 2024. The Inventory is Dark and Full of Misinformation: Understanding Ad Inventory Pooling in the Ad-Tech Supply Chain. In *2024 IEEE Symposium on Security and Privacy (SP)*, 1590–1608.

Vo, Q. M.; Cao, N. T.; and Ton-That, A. H. 2020. Unsafe Image Classification Using Convolutional Neural Network for Brand Safety. In *Proceedings of the Asia-Pacific Conference on Computer Science and Data Engineering*. IEEE.

von Hohenberg, C.; Menchen-Trevino, B.; Casas, E.; and Wojcieszak, M. 2021. A list of over 5000 US news domains and their social media accounts. https://doi.org/10.5281/zenodo.5681483.

Wang, X.; Cheng, F.; Sun, H.; Liu, G.; and Zhou, C. 2018. Adult Image Classification by a Local-Context Aware Network. In *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, and our audit uses only publicly available news pages and seeks to reduce existing harms.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Section 3.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Sections 3, 5, and 6.
   (e) Did you describe the limitations of your work? Yes, see Section 5.
   (f) Did you discuss any potential negative societal impacts of your work? No, though we discuss our recommendations for prior work and industry changes in Sections 5 and 6.
   (g) Did you discuss any potential misuse of your work? No, we do not anticipate how the results of our audit could be used negatively.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we document the software used to collect ratings in the Introduction, and explain how it is used in Section 3.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see Section 3 (though experiments are not ML).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Section 3.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, and all external datasets/tools (e.g., Tranco, U.S. News publisher list) are cited in the References.
   (b) Did you mention the license of the assets? No, because the tool used was open-source.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, a Github repository URL is available for the processed data.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Section 3.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR)? Yes, we intend to make the datasets FAIR, bu making the dataset (raw and processed) and metadata publicly available.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

A Examples of Collected and Processed Data

Figures 6 and 7 provide illustrative examples of the brand safety datasets referenced in our study. Figure 6 displays the original GPT-generated classifications for IAS, while Figure 7 shows the corresponding processed format prepared for analysis.

adt: veryLow
alc: veryLow
dlm: veryLow
drg: veryLow
hat: veryLow
off: veryLow
vio: veryLow
ias-kw: IAS_13436_KW, IAS_3005150_PG, IAS_3006647_PG, IAS_3005110_PG, IAS_3005045_PG, IAS_3006819_PG, IAS_3005044_PG, IAS_3005042_PG, IAS_3005040_PG

Figure 7: Example of Collected GPT Brand Safety Data for IAS

B Frequency of Specific Brand Safety Categories

Given we could not find any previous works that collected brand safety data for news articles, we feel there is some benefit in presenting descriptive results. For example, we showed that articles on breaking/general news websites are more likely to be rated as unsafe than articles from domains which focus on a specific topic (e.g., business). For this same reason, we want to present the specific content categories which are most likely to be rated as unsafe on news articles. These results are not important to our main analysis, thus we present them here in the Appendix.

B.1 Oracle

Oracle has the simplest specific classifications system of the three providers. There are only 10 categories, as shown in Table 1. Classification into each category is only binary, and there are no risk levels. The frequency with which each of these 10 categories is found in the 234 articles classified as unsafe by Oracle are visualized in Figure 9. The "Crime" and "Death & Injury" columns are by far the most common, with 120 and 118 occurrences, respectively.

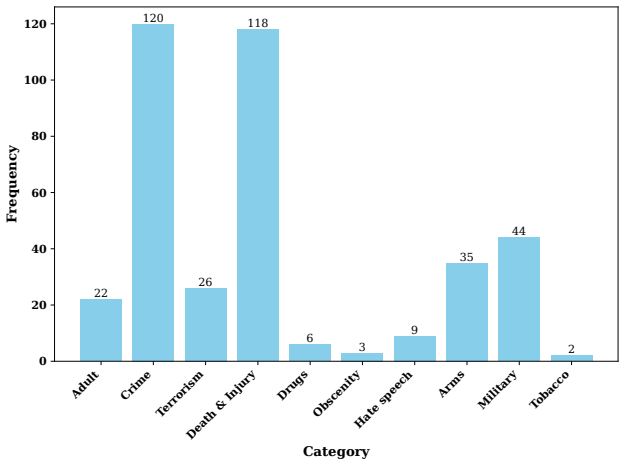

Figure 9: Oracle Specific Classifications

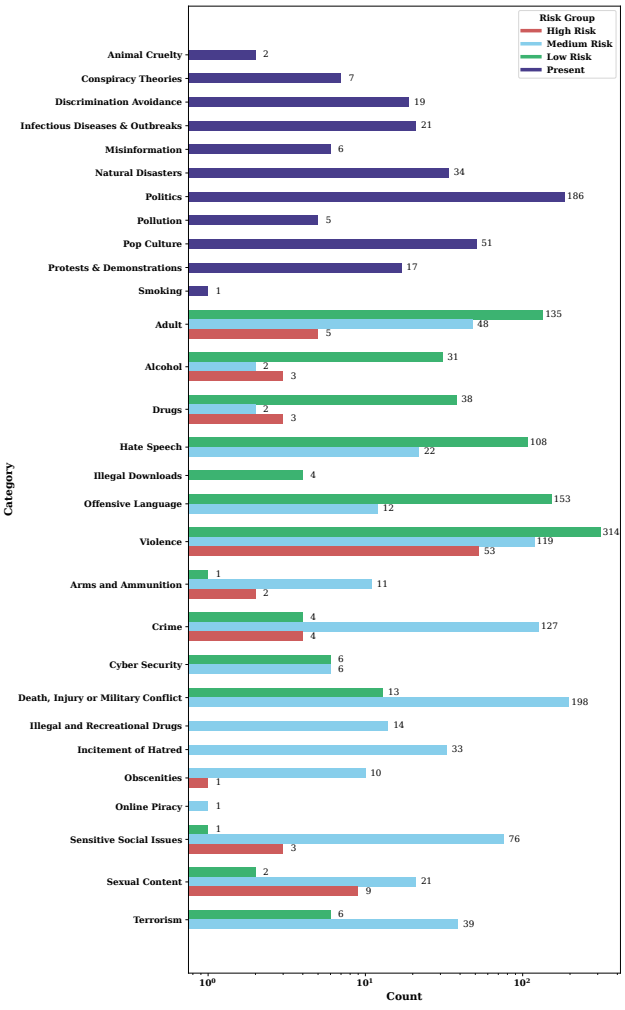

Figure 10: IAS Specific Classifications

MOAT IAS DV

Article ID	Domain URL	m_safety	m_categories	adt	alc	dlm	drg	hat	off	vio	ias-kw	ias_admants	BSC
1	bloomberg.com	unsafe	['gv_death_injury', 'moat_unsafe', 'gv_military']										['84132004', '80000200', ...]
2	forbes.com	safe	['moat_safe']	veryLow	veryLow	low	high	veryLow	veryLow	veryLow	['IAS_3006647_PG', 'IAS_3005123_PG', ...]		['84122001', '80023001', ...]
3	cnn.com			medium	veryLow	veryLow	veryLow	low	veryLow	low	['IAS_13149_KW', 'IAS_1509973_PG', ...]		
4.	theglobeandmail.com	unsafe	['gv_arms', 'moat_unsafe', 'gv_death_injury']									['S_9345', 'S_6900',.....]	

Figure 8: Example of Processed Brand Safety Data

B.2 Integral Ad Science (IAS)

The IAS specific content category classification system is the most complex of the three providers as described in Section 3.3. We show the frequency of all 29 IAS categories in the 895 articles rated as unsafe by IAS in Figure 10. Note that we do not consider the additional "News", "Entertainment", or "Video Gaming" tags and just plot each category according to risk level. We do separately show the frequency of the additional "News", "Entertainment", or "Video Gaming" tags in Figure 11.

Of the binary categories which are only measured for presence, Figure 10 shows that "Politics" is by far the most common with 186 occurrences. For the categories allowing a risk level to be assigned, categories with over 100 total classifications include "Adult", "Hate speech", "Offensive Language", "Violence", "Crime", and "Death, Injury, or Military Conflict". There are some similarities to Oracle, where "Crime" and "Death & Injury" were the two most common categories. However, "Adult" and "Hate Speech" were not very common in the Oracle classifications. Of the 18 categories with a risk level assigned, we find that "High Risk" is rarely used. The one exception is that 53 articles are rated as high risk with respect to the "Violence" category. There are only 30 high risk ratings among the other 17 categories combined.

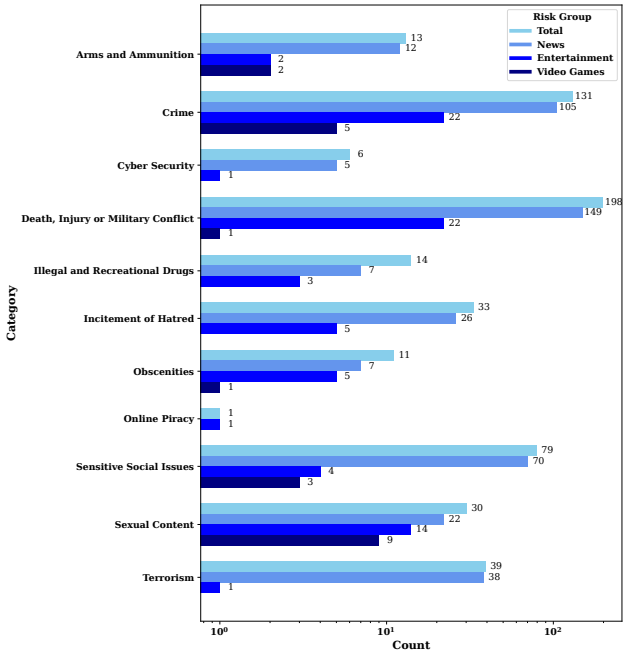

Figure 11: IAS Specific Classifications with News, Entertainment, and Video Gaming Tags

B.3 DoubleVerify (DV)

We collected data on 19 specific categories as described in Section 3.3. We chart the frequency of all 19 categories in Figure 12. Similar to Oracle, the "Crime" and "Death & Injury" are among the most common categories, with over 100 occurrences each. Similar to IAS, the "Violence" category is very common with 178 occurrences. The only other category with over 100 total occurrences is the "Celebrity Gossip" category (which does not exist under the other classification systems). DV rarely rates content as unsafe with "High Risk" as it only occurs 10 times in total (7 for "Adult & Sexual", 2 for "Profanity" and, 1 for "Alcohol"). Oddly, the "Low Risk" level is also not found frequently. It is only used 90 times, which is quite low when considering that the "Medium Risk" level is found 662 times across all categories.

C Appendix: Categorization of Domains

We classify each domain in our sample into a category of news so we can see if patterns exist in the classifications of similar domains. We take a simple approach to classifying each domain. We search for the website on Google and then record the blurb that is generated to describe the website. For each domain, we have copied that text (or a portion of that text) into the "Description" column of Table 3. For a few domains, there is no generated blurb. For these domains, we visit the about page on the website and copy the description from there. These domains are marked with an asterisk in Table 3.

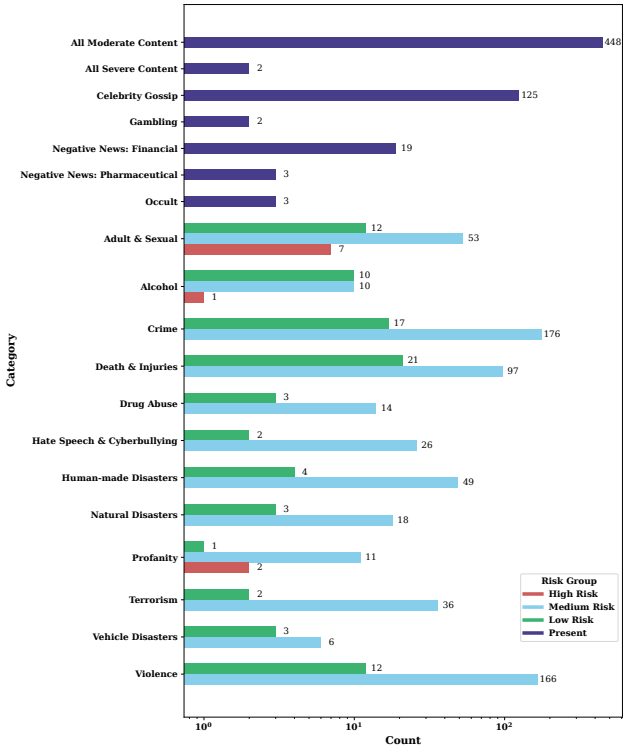

Figure 12: DV Specific Classifications

Once obtaining the descriptions, we can easily classify each domain. Any description that contains the words "breaking... news" is classified into the "Breaking News" category. If the description states it is the online format for a newspaper (e.g., azcentral.com) or if the descriptions state their focus as news for a specific global or local market (e.g., cnn.com or jsonline.com), then we classify them as "General News". Any domain that has a description with a long list of unrelated topics covered (e.g., theatlantic.com or usatoday.com) is also classified as "General News". In the main results section we combine domains which are classified as "Breaking News" or "General News" into one category because we see that the content on these domains is very similar. Domains with a description that clearly states their focus is in one area (e.g., wired.com and jalopnik.com) are classified into a category that is named to align with that focus (e.g., "Technology" and "Cars"). For most of these domains which focus on a specific area, they are in a category of their own. The exceptions are the ten domains in the "Finance/Business" category and the five domains in the "Entertainment" category. We merge the business and finance focused domains because there are several which would fall under both categories anyway (e.g., bloomberg.com and cnbc.com). The domains classified as "Entertainment" either explicitly state that they focus on entertainment news (e.g. variety.com), or we consider their focus to be purely for entertainment (e.g. kotaku.com focuses on video gaming news).

Domain	Description	Categorization
azcentral.com	The digital home of The Arizona Republic newspaper, with breaking news and in-depth coverage of sports, things to do, travel and opinions	Breaking News
barrons.com	Barron's is a leading source of financial news	Finance/Business
bizjournals.com	The Business Journals features local business news from 40-plus cities across the nation	Finance/Business
bloomberg.com	Bloomberg delivers business and markets news, data, analysis, and video to the world	Finance/Business
businessinsider.com	Business Insider tells the global tech, finance, stock market, media, economy, lifestyle, real estate, AI and innovative stories you want to know	Finance/Business
buzzfeed.com	BuzzFeed has breaking news, vital journalism, quizzes, videos, celeb news, Tasty food videos, recipes, DIY hacks, and all the trending buzz	Breaking News
cnbc.com	CNBC is the world leader in business news and real-time financial market coverage	Finance/Business
cnn.com	CNN is the world leader in news and information and seeks to inform, engage and empower the world.	General News
dailymail.co.uk	Get the latest breaking news, showbiz & celebrity photos, sport news & rumours, viral videos and top stories	Breaking News
dailytelegraph.com.au	News and Breaking News Headlines Online including Latest News from Australia and the World.	Breaking News
desmoinesregister.com	The Des Moines Register is the number one source for Des Moines and Iowa breaking, politics, business, agriculture, Iowa sports and entertainment news.	Breaking News
detroitnews.com	Your first source for breaking news, local in-depth reporting, and analysis of events important to Detroit and Michigan	Breaking News
dispatch.com	The Columbus Dispatch is the number one source for Columbus and Ohio breaking politics, business, obituaries, Ohio sports and entertainment news.	Breaking News
economist.com	Get in-depth global news and analysis. Our coverage spans world politics, business, tech, culture and more	General News
euronews.com	European and international latest breaking news, economic news, business news and more.	Breaking News
forbes.com	Forbes is a global media company, focusing on business, investing, technology, entrepreneurship, leadership, and lifestyle.	Finance/Business
ft.com	News, analysis and opinion from the Financial Times on the latest in markets, economics and politics.	Finance/Business
hollywoodreporter.com	Movie news, TV news, awards news, lifestyle news, business news and more	Entertainment
huffpost.com	Read the latest headlines, news stories, and opinion from Politics, Entertainment, Life, Perspectives, and more.	General News
investors.com	Perform stock investment research with our IBD research tools to help investment strategies.	Finance/Business
jalopnik.com*	Jalopnik is a news and opinion website about cars	Cars
jsonline.com	Milwaukee and Wisconsin news, sports, business, opinion, entertainment, lifestyle and investigative reporting	General News
kotaku.com	Gaming Reviews, News, Tips and More.	Entertainment
law.com	The premier global source for trusted and timely legal news, analysis and data.	Legal News
marketwatch.com	MarketWatch provides the latest stock market, financial and business news	Finance/Business
mashable.com	Mashable is a global, multi-platform media and entertainment company.	Entertainment
metro.co.uk	Metro.co.uk: News, Sport, Showbiz, Celebrities from Metro.	General News
msnbc.com	MSNBC breaking news and the latest news for today	Breaking News
news.com.au	The Latest on the news that matters to you - sport, entertainment, finance and politics.	General News

Domain	Description	Categorization
newyorker.com	The New Yorker is an American magazine featuring journalism, commentary, criticism, essays, fiction, satire, cartoons, and poetry.	General News
nymag.com	New York Magazine obsessively chronicles the ideas, people, and cultural events that are forever reshaping our world.	General News
nytimes.com	Live news, investigations, opinion, photos and video by the journalists of The New York Times from more than 150 countries around the world.	General News
qz.com	Quartz is a guide to the new global economy for people who are excited by change. We cover business, finance, economics	Finance/Business
realclearpolitics.com	RealClearPolitics (RCP) is an independent, non-partisan media company that is the trusted source for the best news, analysis and commentary.	Political News
reuters.com	Find latest news from every corner of the globe at Reuters.com, your online source for breaking international news coverage.	Breaking News
telegraph.co.uk	A British daily broadsheet newspaper	General News
theatlantic.com	The Atlantic covers news, politics, culture, technology, health, and more, through its articles, podcasts, videos, and flagship magazine.	General News
thedailybeast.com*	The Daily Beast delivers award-winning original reporting, fact-informed analysis, and sharp opinion in the arena of politics, pop-culture, and power and reaches more than 1 million readers a day.	General News
theglobeandmail.com	The Globe and Mail offers the most authoritative news in Canada, featuring national and international news.	General News
theguardian.com	Latest US news, world news, sports, business, opinion, analysis and reviews from the Guardian, the world's leading liberal voice.	General News
theroot.com*	Black News and Black Views with a Whole Lotta Attitude	General News
thesun.co.uk	Breaking headlines and latest news from the UK and the World	General News
theweek.com	Concise, twice-daily news digests curated by our editors. Distilled from dozens of the world's most trusted news sources	General News
time.com	Breaking news and analysis from TIME.com. Politics, world news, photos, video, tech reviews, health, science and entertainment news.	Breaking News
usatoday.com	Current national and local news, sports, entertainment, finance, technology, and more	General News
vanityfair.com	Vanity Fair is an American monthly magazine of popular culture, fashion, and current affairs	Fashion
variety.com	Entertainment news, film reviews, awards, film festivals, box office, entertainment industry conferences.	Entertainment
vox.com	Vox is a general interest news site for the 21st century.	General News
vulture.com	Vulture is a New York Magazine site providing continuous entertainment news covering TV, movies, music, art, books, theater, comedy, podcasts, celebrities,	Entertainment
washingtonpost.com	Breaking news, live coverage, investigations, analysis, video, photos and opinions	Breaking News
wired.com	Focuses on how emerging technologies affect culture, the economy, and politics	Technology

Table 3: Categorization of Domains
Note: * means no Google blurb was available, so description is from the website itself