

# Differential Tracking Across Topical Webpages of Indian News Media

Yash Vekaria<sup>\*α</sup>, Vibhor Agarwal<sup>\*α</sup>, Pushkal Agarwal<sup>β</sup>, Sangeeta Mahapatra<sup>γ</sup>,  
Sakthi Balan Muthiah<sup>α</sup>, Nishanth Sastry<sup>δ</sup>, Nicolas Kourtellis<sup>ζ</sup>

<sup>α</sup>The LNM Institute of Information Technology, Jaipur, India

<sup>β</sup>King's College London, London, United Kingdom

<sup>γ</sup>German Institute for Global and Area Studies, Hamburg, Germany

<sup>δ</sup>University of Surrey, Surrey, United Kingdom

<sup>ζ</sup>Telefonica Research, Barcelona, Spain

<sup>α</sup>{yash.vekaria.y16, vibhor.agarwal.y16, sakthi.balan}@lnmiit.ac.in, <sup>β</sup>pushkal.agarwal@kcl.ac.uk  
<sup>γ</sup>sangeeta.mahapatra@giga-hamburg.de, <sup>δ</sup>n.sastry@surrey.ac.uk, <sup>ζ</sup>nicolas.kourtellis@telefonica.com

## ABSTRACT

Online user privacy and tracking have been extensively studied in recent years, especially due to privacy and personal data-related legislations in the EU and the USA, such as the General Data Protection Regulation, ePrivacy Regulation, and California Consumer Privacy Act. Research has revealed novel tracking and personal identifiable information leakage methods that first- and third-parties employ on websites around the world, as well as the intensity of tracking performed on such websites. However, for the sake of scaling to cover a large portion of the Web, most past studies focused on homepages of websites, and did not look deeper into the tracking practices on their topical subpages. The majority of studies focused on the Global North markets such as the EU and the USA. Large markets such as India, which covers 20% of the world population and has no explicit privacy laws, have not been studied in this regard.

We aim to address these gaps and focus on the following research questions: Is tracking on topical subpages of Indian news websites different from their homepage? Do third-party trackers prefer to track specific topics? How does this preference compare to the similarity of content shown on these topical subpages? To answer these questions, we propose a novel method for automatic extraction and categorization of Indian news topical subpages based on the details in their URLs. We study the identified topical subpages and compare them with their homepages with respect to the intensity of cookie injection and third-party embeddedness and type. We find differential user tracking among subpages, and between subpages and homepages. We also find a preferential attachment of third-party trackers to specific topics. Also, embedded third-parties tend to track specific subpages simultaneously, revealing possible user profiling in action.

## 1 INTRODUCTION

India has seen exponential growth in online users with the recent advancements in network coverage and cheap internet rates. This enormous growth coupled with the lack of a data privacy law in India<sup>2</sup> potentially allows for unregulated tracking of users. Though

tracking has its advantages, it also gives opportunities to the third-parties to track the readers and build their profiles, which can be used further for targeted advertisements and influence campaigns. Studies have shown that news websites have more tracking than other categories [2, 8]. Therefore, we choose to focus on Indian news websites to demonstrate tracking in this under-studied country, which happens to be the second largest internet market in the world [16].

An important observation of recent studies [5, 27] is that researchers have been focusing on the homepages of websites for understanding the tracking ecosystem around the world due to its ease in scaling the approach. Our focus on one country (India) and one type of websites (news), allows us to dig deeper to understand how tracking varies across topical subpages of a website. We believe this is an orthogonal, yet important contribution that needs to be replicated in other, better studied settings such as the USA and the European Union. News is ideally suited for such a study across different subpages of a website since different users are likely to be interested in different topics.

To this end, we undertake, to our knowledge, the first in-depth study of Web tracking in more than 100 top Indian news websites<sup>3</sup>. In this work, we address three research questions: (1) Is there any difference in tracking on topical subpages of Indian news websites and their homepages? (2) Do third-party trackers have preferential attachment to specific topical subpages? (3) Does this preference have any connection with the similarity of content shown to users on topical subpages? Towards addressing these questions, we propose DiBETS, a method for classifying topical categories based on keywords of their URLs, and demonstrate differences in tracking of 112 homepages of websites and across 1.3K topical subpages.

We find that 11 out of 15 topical subpages are *tracked more intensely* than their homepages – topical subpages can have as high as 44% more cookies. For some topical subpages, there are additional third-parties (TP), which are involved in more fingerprinting and analytics when compared to their homepages. We also find that there is a *specialisation* of third-party partners, who prefer to advertise or track in specific topics. Some of these appear to be aligned to specific business affinities of the trackers for different topics. For

\*Both the authors contributed equally to this research.

<sup>2</sup>Personal Data Protection Bill, tabled in Indian Parliament in December 2019, is still with the Joint Parliament Committee for review.

<sup>3</sup>Our dataset and code is released via this link for non-commercial research usage: <http://tiny.cc/india-topic>.

example, music websites, which track or advertise on the ‘entertainment’ subpage, or games and sports websites advertising on ‘sports’ subpage. We also find some counter-intuitive preferences, such as a paint company tracking in Coronavirus-related subpages. We find that such aberrations can be explained in some cases (in this case, the paint company recently claimed to have released an *anti-Covid-19 paint*<sup>4</sup> and also introduced sanitisers<sup>5</sup> in its Health and Hygiene segment during the start of the pandemic). Using clustering approaches, we show that embedded TP trackers tend to track specific subpages simultaneously, revealing possible user profiling in action. Such preferential attachments may be among smaller players who have affinities towards specific topics and subpages. However, major TP players (such as `doubleclick.net`) are present on all topics. Finally, we note that privacy policies are nearly always in English, even on regional Indian language news websites, which make it hard for the audience to truly understand or consent to ad-related privacy policies.

## 2 RELATED WORK AND BACKGROUND

Previous studies focused on tracking in different categories of websites (e.g., News [8, 27], Adult [28], Education, etc.) and types of users. In [2] authors showed that tracking on various demographics, simulated by artificial user personas using categorical websites list from Alexa Inc. [4], can significantly differ. For instance, after visiting around 20 websites related to youth and women, there were about 60 and 300 distinct TP trackers, respectively, in an artificial user’s history. Among all personas, youth demographics had the least number of trackers, possibly because of strict laws such as USA’s Children Online Privacy Protection Rule [6], which protects the youth population from being tracked. In [27] authors used McAfee SmartFilter service [20] to obtain categories of one million top trending websites. They found that tracking on news websites was more than other types [8, 27].

Interestingly, internal page discovery and analysis have only been recently studied [5, 27]. Subpages of a website cover various different topics, and can be based on user preferences including trends and themes of the website [5], as well as catering to a user’s interest in ‘hard’ or ‘soft news’ with their differing content and style of presentation [23]. These works emphasize the need for deeper analysis of internal webpages, which differ in content, performance, and user tracking from the landing/home page. In [5], the landing page of a website was compared with several internal pages of the same website in order to analyse the difference in the content and the performance from the landing page. This was done for 20K pages from about 1K websites. Their main contribution was to show that the Web performance measurement and optimization need not be only based on the landing pages but also on the internal pages. In our work, the main objective is to look at the tracking ecosystem in the topical subpages of news websites. Our focus is more related to [27], where they analyse the relationship between third-parties in over 10K websites. They build third-party trees that capture the loading sequences and the dependencies of various third-parties. Their study indicates that there are substantially more cookies in the subpages

<sup>4</sup><https://realtypusmag.com/asian-paints-launches-anti-covid-19-paint/>

<sup>5</sup><https://english.jagran.com/business/exclusive-asian-paints-fights-covid-crisis-with-safe-painting-and-san-assure-services-focuses-on-health-and-hygiene-products-10014656>

than the site’s landing page (36% more), which implies that studies based only on the landing pages could be biased.

There are various studies which use different computational methods to quantify and understand topics on subpages, such as Word2Vec in the Indian context for Tamil language [22], Naïve Bayes in Arabic language [11], etc. These topic identification techniques can help publishers in providing personalised news and understanding the emergence of new topics over time [7, 15, 29]. In fact, previous studies have focused on personalised news topic selection based on user profiles and browsing history, for e.g., using click behaviour [19], and public news selection through raw text analysis [26]. The present study is different as it is not about topic selection or salience from end user perspective but one derived from the presence of Web trackers on topical subpages. We study which topics attract most trackers for user targeting. Furthermore, news topic analyses are usually related to media slant, i.e., how news publications tilt towards ideological positions based on readers preference, ownership consideration, and market compulsions [12], or agenda setting function of news media, i.e., how mass media influences public opinion on certain issues/agendas [24]. Our study on tracking of topical subpages can aid these studies as trackers can provide detailed information on the profiles of the users and their associated browsing behaviour, allowing for targeted ads and slanted news coverage targeted at users.

Indeed, while topical tracking has been studied across websites, tracking on topical subpages within websites is still fairly unexplored. Hence, we build upon past studies, to present topical tracking in Indian news websites. We focus on Indian news as it has huge amount of trackers compared to the general Web. A reason can be that India witnessed massive growth in Internet users in the past few years [3]. In general, such Indian news websites have different categories of news in their homepage, such as International, National, Politics, Business, Sports, etc. The collection of data from homepages and their subpages in automated way is prone to encounter problems since the topical names might be different in various websites. For example, in the Times of India, there is a category called *India*, but in The Hindu, the same semantically-related category is called *National*. Moreover, there are new trending topics in most Indian newspapers like *Coronavirus/Covid*, and there are examples of topics such as *Manoranjan*, which is a Hindi word for ‘entertainment’. So, in order to collect the data and segregate it into categories, we built a new model called *DiBETS* (Dictionary-Based Extractor of Topical Subpages) which classifies the subpage URLs into different categories. This supplements more manual and qualitative approaches previously undertaken in the field [21]. An additional differentiation from such previous works is that they do not focus on the Indian News landscape, which is very large and highly diverse.

## 3 METHODOLOGY AND DATASET COLLECTION

In this section, we describe our methodology to study tracking in topical subpages. We first provide details on how we identify topical subpages per website using DiBETS (Dictionary-Based Extractor of Topical Subpages), our novel method (Sec. 3.1). Next, we discuss application of DiBETS to Indian news context (Sec. 3.2) and outline our tracking and content crawling of webpages (Sec. 3.3).

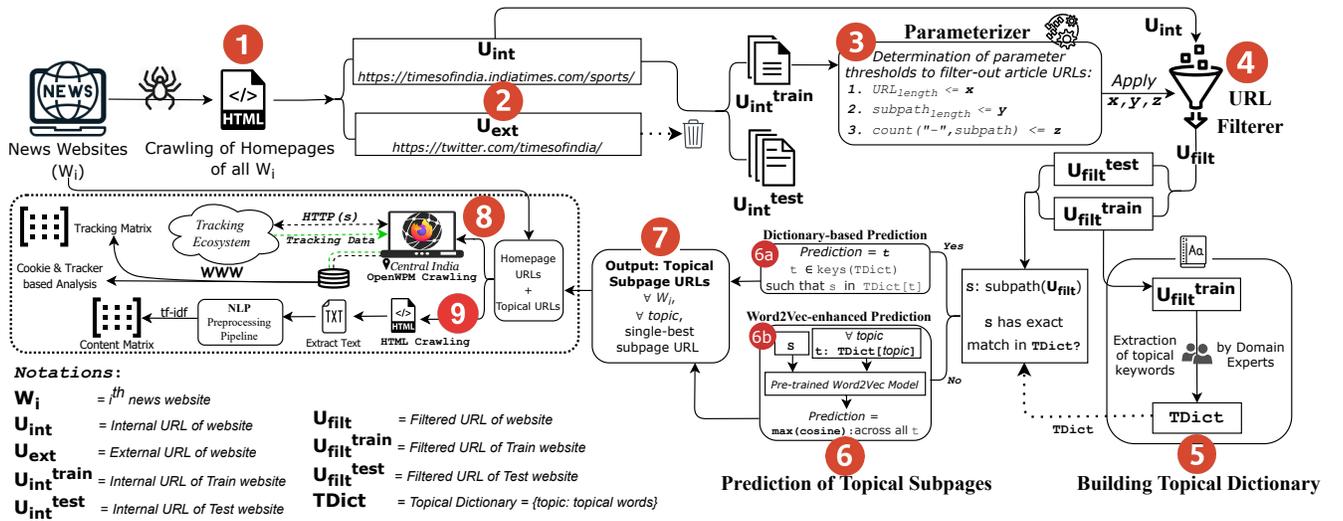


Figure 1: Our methodology to study tracking of topical subpages in news websites.

### 3.1 Automated Extraction of Topical Subpages with DiBETS

We propose DiBETS, a Word2Vec-enhanced method to extract the main topical subpages from a website’s homepage. Since our study focuses on news websites, we tune DiBETS to detect the subpages of 15 standard topics covered on news websites. We choose these topics from a study by Reuters [21], where it compares the focus of 12 topics in the news websites of eleven countries. We regroup some topics used by [21] taking a higher level perspective (e.g., *Business and Finance* and *Economics* those being related, are grouped into one; *Fun/Weird News* is included under *Entertainment*). We also add few other categories: e.g., *Multimedia* is more usefully seen as a separate topic based on the prominent presence of multimedia-based subpages on Indian news websites. In the context of Covid-19, numerous news websites are seen to incorporate a separate subpage for *Coronavirus*; hence, we define it as a separate topic. To study the privacy pages and cookie policy in Indian context, we also add a topic of *Privacy Policy*. Thus, we obtain a master list of 15 topics. We classify a subpage that belongs to topics outside these 15 topics as *Other*. Figure 1 illustrates our methodology, including DiBETS’ main Steps 1-7, and the subsequent data crawling Steps 8-9.

#### 3.1.1 Website Crawling and Extraction of Internal URLs.

DiBETS takes a news website ( $W_i$ ) as an input and first crawls the raw HTML of the website’s homepage (Step 1). In step 2, from the raw HTML of the homepage, all valid embedded URLs are extracted (HTML “href” tags). These URLs are segregated into internal and external URLs. Internal URLs are the ones belonging to the same domain as the homepage. External URLs point to a different domain and are discarded. These two steps are performed collectively for all websites of interest. We select a small portion of websites (and its URLs) to train and tune DiBETS in a supervised fashion, and then apply it on the remaining websites.

**3.1.2 URL Parameterizer and Filterer.** Internal URLs may consist of article URLs or other subpage URLs. Since our objective is to obtain the subpage URLs of different topics focused on the homepage of a website, we filter out article URLs. Manual inspection of a subset of the training set revealed that article URLs are typically longer with respect to number of characters and subpaths (subpath is a portion of URL path between two consecutive “/”, each organized as hyphen-separated words) and have more hyphens than subpage URLs. In fact, these three parameters (*URL length*, *subpath length* of subpaths within that URL, and *hyphen count* within a subpath) have clear bi-modal distribution, easily separating the two types of URLs (article vs. subpage URLs). Thus, we identify automatically a threshold for each parameter (i.e.,  $x$ ,  $y$ , and  $z$  respectively for  $length(URL)$ ,  $length(subpath)$  and  $count("-", subpath)$ ), as the point between the two modes in the frequency distribution of each parameter. Step 4 applies these thresholds and keeps as subpage URLs the internal URLs that have, for all three parameters, numbers that are less than or equal to the thresholds found.

**3.1.3 Topical Dictionary.** In Step 5, we build the Topical Dictionary by adapting one-time manual process with the help of a domain expert who is well-versed with the thematic structure followed by news websites. The dictionary is built with the remaining subpage URLs of the train set from Step 4. For each URL, the expert attributes the URL to one of the 16 (15 topical subpages plus “Other”) by either looking at the URL or actually visiting the URL and observing its content. Additionally, the expert associates keywords present in the URL’s subpaths that can be suggestive of the subpage’s topic. Following this, we generate a dictionary ( $TDict$ ) consisting of ( $key, value$ ) pairs, where key is any of the 16 topics, and value is a list of keywords found in the subpage URLs of the said topic.

**3.1.4 Prediction of Topical Subpages.** To determine the topic of a URL in Step 6, the subpaths of each filtered URL are considered hierarchically from high-level subpath to a low-level, one by one,

and checked in the Topical Dictionary. If the subpath is a generic phrase (e.g., “category” or “section”), we continue to the next one.

**Dictionary-based Prediction (Step 6a):** If there is an exact match of the subpath with any keyword in the Topical Dictionary, then the topic (key) of that keyword is assigned to that subpage URL. Else, we use *Step 6b* to get a topic.

**Word2Vec-enhanced Prediction (Step 6b):** DiBETS uses a pre-trained Word2Vec model on English CoNLL17 corpus with a vocabulary size of 4027169 words [18]. This model is trained using a Word2Vec Continuous Skip-gram algorithm with a vector size 100 and a window of 10 words.

To predict the topic of a subpage URL, first, we tokenize the subpath, remove stop words, and compute a combined word embedding  $comb\_emb[subpath]$  for all tokens found. Similarly, for each topic  $T_i$ , we also pre-compute a combined word embedding  $comb\_emb[T_i]$  by summing up word embedding of all its keywords in  $TDict$ . Finally, we compute the cosine similarity between these two embeddings. The predicted topic is determined as per Eq (ii), if the corresponding maximum value of cosine similarity is greater than the threshold value. The same automated method from Sec. 3.1.2 is used to set the threshold between the bi-modal histogram of maximum cosine score amongst the cosine scores of each internal URL with all the topics.

$$T_{pred} = \arg \max_{T_i \in keys(TDict)} \text{cosine}(comb\_emb[subpath], comb\_emb[T_i]) \quad (ii)$$

**3.1.5 Determining Single Best subpage URL for each topic.** Each filtered URL is assigned one of the 16 topics by DiBETS. As a result, there may exist more than one subpage URL (of the same website) for a given topic in the output of *Step 6*<sup>6</sup>. So in the final stage (*step 7*), the model determines and outputs a single best subpage URL for each topic per website (depending upon the topics that are detected by the model in that website). Out of all the URLs assigned to the same topic  $T$ , DiBETS weighs each URL by the following ratio<sup>7</sup>:

$$weight = \frac{\text{cosine}(comb\_emb[all\ subpaths], comb\_emb[TDict[T]])}{count(tokens\ across\ all\ subpaths)} \quad (iii)$$

These URLs are then ordered in descending order of the weight Eq (iii) and checked from the top if the URL’s highest-level subpath is entirely present in our Topical Dictionary (corresponding to the assigned topic  $T$ ) or not. If yes, DiBETS outputs it as the main subpage URL of  $T$  for a website. Otherwise, it returns the first URL in the above ordering.

## 3.2 DiBETS application to Indian News Context

The main input to our DiBETS model is the raw HTML of website homepages. We used a dataset we collected previously, of 123 popular Indian news websites [3]. Using an automated Python script, we performed two separate crawls of the homepages of all 123 websites

<sup>6</sup>E.g. <https://www.xyz.com/sports/>, <https://www.xyz.com/sports/cricket/> and <https://www.xyz.com/sports/athletics/> are all assigned to topic “Sports”.

<sup>7</sup>Shorter the URL, higher the probability of it being the main subpage URL. This is because shorter URLs have single or smaller hierarchy of subpaths and represents the main subpage for a topic. Longer URLs add more hierarchy to subpaths, resulting in sub-subpage URLs. Hence, the ratio is inversely weighted by the number of tokens in subpath after stop-word removal.

in July and December 2020. We compared the output of DiBETS across both crawls and the results of topical subpages were stable across time and topics (fraction of websites of topic  $i$  in July and December remained almost constant, with a difference of  $\sim 4\%$  (min=0.98%, max=22%)). Our analysis in Sec. 4 is on the December 2020 crawl.

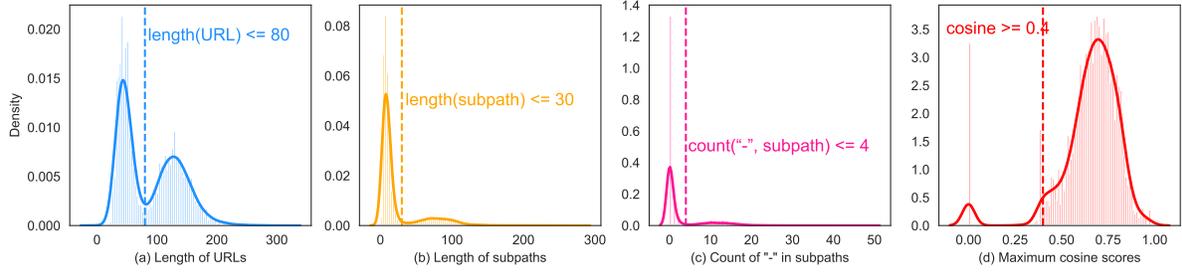
After passing the 123 websites to *Step 1* of DiBETS, we could retrieve the HTML of only 112 websites (11 websites forbade access). In them, we found 15540 internal and 2718 external URLs. Then, we identified the following as thresholds for the three parameters involved:  $length(URL) \leq 80$ ,  $length(subpath) \leq 30$ , and  $count("-", subpath) \leq 4$ . Also, the maximum cosine score threshold in the Word2Vec (Step 6b) was found to be 0.4 (also shown in Figure 2). The Topical dictionary built comprises a total of 1491 distinct keywords. Thus, as the final output of DiBETS in *Step 7*, we obtained 1275 subpage URLs for different topics detected across sites.

**Evaluation of DiBETS Performance.** We asked two domain experts from political science and journalism backgrounds to verify the prediction of the best topical subpages across 112 websites by DiBETS. Each expert was presented with the model’s output of a randomly selected set of 250 subpage URLs alongside the topic predicted by the model. Then, they were independently asked to make a binary decision whether the model selection for best topical subpage was correct or not. We computed 96% agreement between experts, and an inter-annotator agreement with Cohen’s Kappa score of 0.66, representing a “substantial agreement”.

**Generalisability of DiBETS.** To test the generalisability of our model, we also applied DiBETS in the USA context on a set of top 15 American news websites (based on Alexa ranks and Facebook followers), crawled in January 2021. We applied similar process for training and testing the method and asked our experts to perform annotation for these results as well. We observed an agreement of 95.67% and Cohen’s Kappa as 0.72. These results give us a confidence that DiBETS is applicable in other contexts outside India.

## 3.3 Website Crawling for Tracking and Content

As depicted in Step 8 of Figure 1, we crawled websites and collected traffic data using OpenWPM [8, 9], a web privacy measurement framework. We used most settings at default, with Firefox browser and third-party tracking enabled and tracking protection as false so that nothing is blocked. We also enabled `http_instrument` which logs HTTP(s) responses, requests and redirects, using a selenium headless browser to perform crawling and setting `javascript_instrument`, and `javascript_cookies` to true to store javascript cookies information. We performed five stateless crawls of the  $112+1275=1387$  URLs (output from DiBETS), consisting of news websites’ homepages and topical subpages during 1-31 December 2020. In total, we logged more than 1.2M cookies, 1.7M HTTP(s) requests, 1.5M responses, and 297K redirects. Since stateless crawls make each website visit independent, parallel browser instances were launched to allow simultaneous crawls of these news websites from Central India. Moreover, we performed such crawls with a gap of six days to log the dynamic traffic data and account for infrequent but unavoidable network errors during each crawl.



**Figure 2: Histogram distributions of URL-specific parameters: (a) Length(URL), (b) Length(subpath), (c) Count('-', subpath), (d) Histogram of Word2Vec-based maximum cosine similarity scores of all the internal URLs. Finalized thresholds for each distribution are marked by vertical dashed lines.**

Using the above data, we mapped the presence of 561 distinct third-party trackers found across homepages and topical subpages as a  $16 \times 561$  binary matrix, which we call *tracking matrix*. The binary value signifies whether that third-party exists in at least one topical subpage of the given topic or not.

As we also want to understand the semantic relation between topical subpages across websites, we applied content analysis (Step 9). First, we crawled the HTML of the 1387 URLs, obtaining the HTML source for 1349. Then, we extracted the textual content from each and applied an NLP Pre-processing pipeline (case-folding, eliminating blank spaces, word segmentation, un-accenting words, removing punctuation, lemmatization and stop-word removal). Then, we computed a *tf-idf* vector for each topic, by combining all extracted text from each subpage into one document per topic, thus, constructing a  $16 \times 25168$  matrix with 16 topics and 25168 features. We refer to this as *Content Matrix*. We use both aforementioned matrices in Section 4.5.

## 4 ANALYSIS OF TRACKING ON TOPICAL SUBPAGES

In this section, we analyse and compare the tracking found on Indian news homepages and their topical subpages, in order to find differences between them (Sec. 4.1), types of third-parties involved (Sec. 4.2), dominant tracking entities (Sec. 4.3), preference of them to track specific topical subpages (Sec. 4.4) and how this compares with content (Sec. 4.5). Where appropriate, we use the Kolmogorov-Smirnov (KS) test to assess statistical significance of distribution differences.

### 4.1 Is tracking on topical subpages different from homepage?

To understand the difference between homepage and topical subpages, we first present the existence of topical subpages across news websites in our dataset. Second, we use the number of cookies injected on these webpages as a proxy for tracking as previous studies [3, 8]. Figure 3 (Left) shows the percentage of Indian news websites covered by each topic. Clearly, not all topics are covered by all news websites. *Privacy Policy* appears in 92% of websites, perhaps as an attempt to comply with EU/USA laws. *Multimedia* is in 86% of websites, as users may be interested in news with media content

like videos. We also see ‘hot’ topical pages like *Coronavirus*, mainstream topics like *Politics*, and city-related topic like *Local*, which are covered by nearly one in every two news websites.

Figure 3 (Right) shows the box plot of number of cookies for each topical subpage, ordered by their median number of cookies. We also include their homepage for comparison. We find that 11 out of 15 topics have higher median cookies than the Homepage, indicating more intense tracking on these topical subpages. Interestingly, less frequent topics (occurring in half of websites) such as *Regional* and *Coronavirus* have the highest and second highest number of median cookies, respectively. Also, *Coronavirus* has the highest mean number of cookies (243), which can have up to 1280 maximum cookies for a single webpage on this topic. *Privacy Policy* has the least median number of cookies (28). Interestingly, eight websites do not have privacy policy pages. The remaining 104 websites have one but in English only<sup>8</sup>. In fact, in our dataset, 23 out of 104 websites are in Hindi or other regional languages, but they all have privacy policy pages in English. This suggests that privacy policy subpages in news websites are mostly acting as placeholders and do not attract many users or trackers [17].

**Findings:** Nearly 75% of topical subpages track users more intensely than Homepages. In some cases, this tracking using cookies can be at up to 1.4x than Homepage. Topics *Regional* and *Coronavirus* have the highest and second highest number of cookies, respectively, and have statistically significant more tracking than Homepages with KS-value 0.33 (p-value: 0.05) and 0.18 (p-value: 0.07), respectively. *Privacy Policy* topic tracks users the least with 65% fewer cookies than Homepages. Also, eight websites do not have privacy policy page and wherever it exists, it is in English.

### 4.2 What are the categories of third-party cookies tracking different topical subpages?

Next, we study the type of cookies set additionally to Homepage. We breakdown the cookies into types using domains list by *Disconnect.me* (DL) [14]. This list provides categories of third-party (TPs) like *Advertising*, *Content & Social*, *Analytics* and *Fingerprinting*. Figure 4 (Left) shows the bar plot of topic-wise cookie categorization for all news websites (represented via solid bars) and top ten popular (top as per Facebook followers) websites (represented via

<sup>8</sup>We detect the language of topical pages using ‘Langdetect’ which is built upon Google’s language-detection <https://pypi.org/project/langdetect/>.

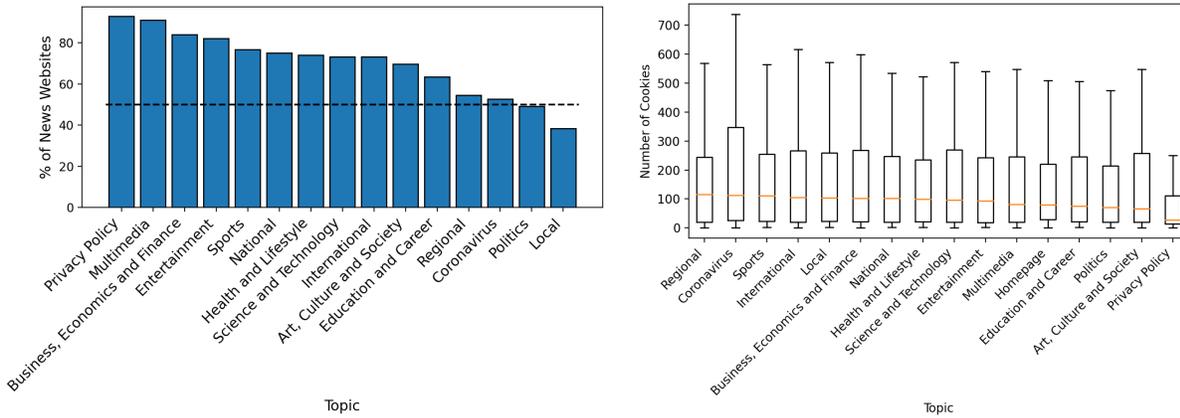


Figure 3: Left: Portion of websites covered by each topic. Right: Number of cookies set in topical subpages across websites.

dashed bars). Topics are sorted in descending order of number of distinct third-parties in all webpages. Note that *Unknown* represents the number of TPs that cannot be categorized using DL. More than 80% of TP domains categorized by DL belong to *Advertising*. This is expected since most news websites generate income through display ads. Interestingly, the second largest portion of TPs belongs to *Fingerprinting*. In the case of all websites (solid bars), *Homepage* has the highest number of distinct TPs followed by other topics. But in the case of top ten websites, it is not so: nine out of 16 topics have more distinct TPs than *Homepage*, with *Health and Lifestyle* having the highest number with KS-value 0.8 (p-value: 0.04). This shows that, in general, more tracking is performed by more distinct TPs on the topical subpages. Surprisingly, we find that 182 *Advertising* and other TP domains are tracking users on *Privacy Policy* pages. The *Unknown* TPs are mostly India-specific, but few belong to other countries as well: e.g., *audiencemanager.de* (Germany), *gsspat.jp* and *fw-ad.jp* (Japan) etc. Majority of the TPs are the ones that have preferential attachment towards some topics like *asianpaints.com* for *Coronavirus*, *gamepix.com* for *Sports* etc. Interestingly, two TP domain names (*s3xified.com* and *xvidoes.com*) appear to be related to Adult websites. We also notice that for top ten news websites, the amount of *Unknown* is less when compared to all news websites.

Figure 4 (Right) shows the bar plot of percentage difference between *Homepage* and subpages of each type of third-party. We find that ten out of 15 topics, such as *Multimedia*, *Health & Lifestyle*, and *Coronavirus*, have more *Fingerprinting* TP domains than the *Homepage*. Moreover, *Science & Technology* and *Art, Culture & Society* also have more *Content & Social* TPs than the *Homepage*.

**Findings:** More than 80% of cookies set on news websites, as categorized by DL, are *Advertising*. The fraction of *Unknown* cookies is less in top ten popular websites, as they probably employ more well known and categorized TP trackers. *Advertising* and other TPs track users on *Privacy Policy* pages as well. In case of top ten news websites, the majority of topics have more distinct TPs than *Homepage*, with *Health & Lifestyle* having the highest number with KS-value 0.8 (p-value: 0.04). In general, topics have higher *Advertising* and *Fingerprinting* cookies than *Homepages* with KS-value 0.9 (p-value: 0.05) and 0.67 (p-value: 0.04) respectively. Among the TPs with *Unknown* category, few domain names are related to Adult-content

websites, while some TPs belong to other countries like Japan, Germany etc.

### 4.3 Are top third-party trackers present in all the places?

In total, 523 distinct TP domains are found setting cookies on news websites and their topical subpages. The top 20 TPs are almost common in all topics. Due to space, we focus on the top ten TPs. In Figure 5 (Left) shows the heat map of percentage of topic subpages in which the top ten TPs appear. Like past studies that have found a handful of top TP trackers present in most websites [2, 3, 8], we also find that the top ten TP trackers are in large fraction of the topical subpages and *Homepages* in our dataset. Interestingly, all pages are found to have some of the top ten TPs with little difference in the TP ranking (based on the number of cookies set). The top TP *doubleclick.net* is present in more than 75% of webpages for all topics except *Privacy Policy*. From rank 3 onward (i.e., *pubmatic.com*), we see a sudden drop in and then fairly consistent amount of websites coverage (40% to 50%). As reported in Sec 4.1, *Privacy Policy* has the lowest number of median cookies, and even the top ten TPs are least interested, as they cover the lowest portion of *Privacy Policy* pages.

**Findings:** A handful of top third-parties track users not only on *Homepages*, but follow them on specific topical subpages as well. Top 20 TPs are almost common in all topics studied. The top TP *doubleclick.net* is present in more than 75% of webpages for all topics except *Privacy Policy*.

### 4.4 Are third-party trackers preferentially attached to specific topical subpages?

As we saw earlier, 11 out of 15 topics have higher median cookies than the *Homepage*, but TPs are not evenly present in all topics. Thus, here we study *which TPs track what* specific topical subpages and if there is preferential attachment of TPs towards topics. Figure 5 (Right) shows the heat map of 12 TPs preferentially attached (100% presence) to specific topics. The '1' in a cell represents that a TP is present only in that specific topic. Overall, we find 12% (64 in

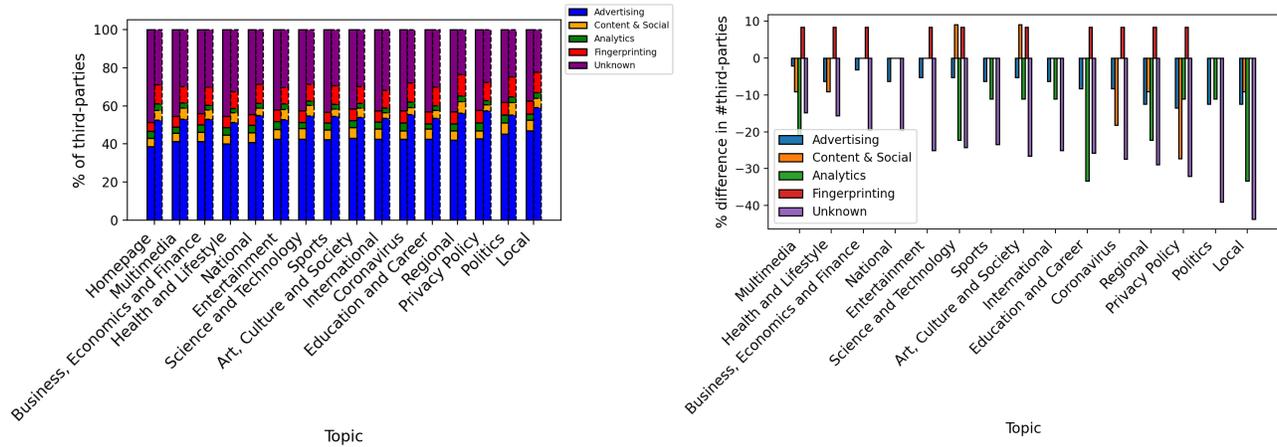


Figure 4: Left: Topic-wise cookie categorization using Disconnect List for all news websites (solid bars) and top 10 news websites (dashed bars). Right: Topic-wise percentage difference in number of third-parties for each category w.r.t. Homepage.

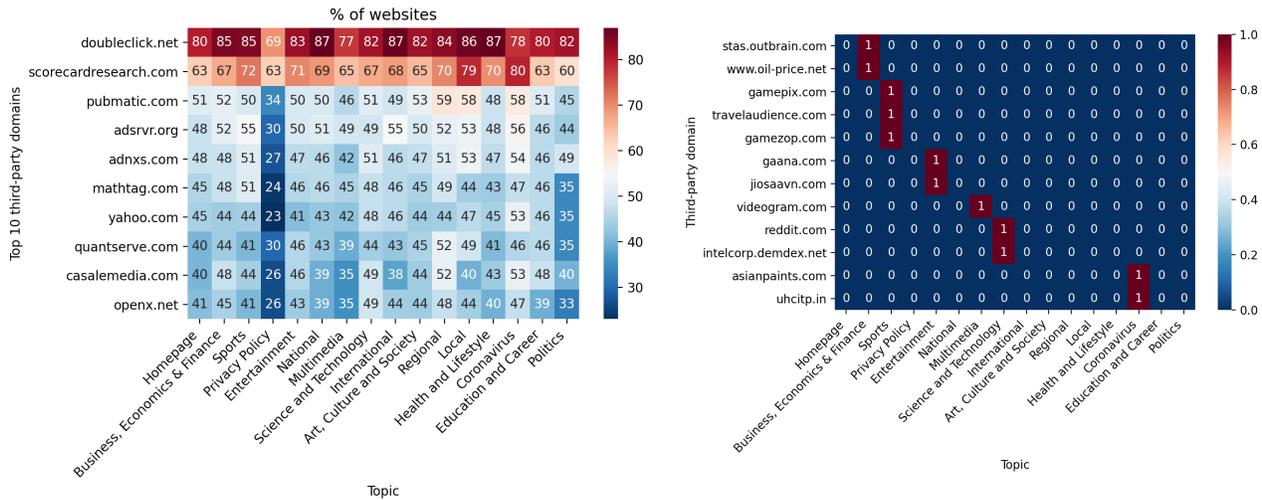


Figure 5: Left: Heat map of top 10 third-parties. Right: Heat map of top 12 preferentially attached third-parties to topics.

523) TPs are preferentially attached to just one of the 15 topics. Few interesting examples are described below:

- Entertainment: *gaana.com* and *jiosaavn.com* are music websites/TPs and, with 3 other TPs, are present only here.
- Sports: *gamepix.com* and *gamezop.com*, with 6 others, are present only in this topic.
- Coronavirus: *asianpaints.com*, which recently released an *Anti-COVID-19* paint, is found only in this topic. Also, *uhcftp.in*, a Universal Health Coverage website, is tracking Coronavirus topic only. 10 other TPs are also present.
- Business, Economics, & Finance: *oil-price.net*, a website for oil prices, with 4 others, is present only in this topic.
- Multimedia: *videogram.com*, a video analytics website boosting video revenue, with 4 others, is present here.
- Science & Technology: *intelcorp.demdex.net*, a technology company, is specifically tracking this topic. *reddit.com* is also present with 2 other TPs.

**Findings:** Specialized third-parties appear to track specific subpages. In fact, there is (100%) preferential attachment of 12% third-parties towards specific topics. This is due to third-parties tracking users' interest on specific news topics, thus, trying to optimize their own marketing and ad-targeting campaigns, or selling user profiling to other TPs not present.

#### 4.5 Unsupervised Clustering of Topics: Do topical subpages show relatedness based on the hosted content or third-party trackers?

While exploring topical tracking, it becomes important to study the two complementary aspects of this ecosystem:

**Users:** The actual content hosted by different topical subpages, and consumed by users. Our aim is to understand if there are semantically similar topical groupings among the subpages.

For this, we analyze the 16 x 25168 tf-idf content matrix (Sec. 3.3).

TPs: The third-parties are embedded in these subpages, and are silently tracking these users. We try to understand if TPs are tracking users of specific profiles, i.e., users that are interested in specific combinations of topics. For this, we analyze the 16 x 651 TP tracking matrix (Sec. 3.3).

We explore such patterns by performing unsupervised K-means clustering on the content and tracking matrices. To reduce noise due to high feature dimensionality, we map the features of both matrices to a lower dimension “N” using PCA. We observed that PCA-reduced  $N = 15$  features cover 100% variance of the original data in both matrices. Thus, we perform K-means clustering for different combinations of (N, K), with  $N, K \in [2, 15]$ . To guide our selection for an appropriate N and K from all combinations, we compute three standard clustering quality metrics [25, 30]: SSE (elbow curve), silhouette coefficient, and gap statistic. Figures 6a and 6b show these metrics across K, for the most appropriate value of  $N = 2$ , for both content- and tracking-based clustering, respectively. As evident, for content clustering (Fig. 6a), we obtain the highest silhouette score for  $N = 2$  and  $K = 4$ . Silhouette score indicates the “tightness” of clusters [25]. Additionally, elbow also occurs at  $K = 4$ . Similarly, among all combinations of N and K, the highest silhouette coefficient for tracking-based clustering (Fig. 6b) occurs at  $N = 2$  and  $K = 2$ . However,  $K = 2$  has a higher SSE and hence we choose  $N = 2$  and  $K = 3$  for clustering.  $K = 3$  is also optimal as per the elbow curve. Overall,  $N = 2$  preserves  $\approx 30\%$  variance for both content and tracking matrices.

Figure 6c shows the resultant clusters obtained by performing K-means with  $K = 4$  on content. Many of the topics are clustered together, revealing a possible similar use of language across the board. Content in *Entertainment* and *Multimedia* subpages, although different from other topics, is very similar to each other. Finally, *Art, Culture, and Society*, as well as *International* seem to be prominently different from the content of other topics, and are assigned to their own cluster. Figure 6d depicts the other side: clusters based on TP tracking. We observe that *Local, Regional* and *Politics* have been tracked differently to other topics, but similarly with each other. This means same trackers are interested in users viewing these topics together. This can be due to the local and regional political elections happening in different parts of India during the time of crawling, which could have attracted similar trackers. Interestingly, *Coronavirus*, which is a transient trend, is tracked most differently from the conventional topics. This is probably because there are a lot of new and unexpectedly higher number of trackers investing on topical subpages of *Coronavirus* than others, since it has become a hot topic. Also, Homepage, although it is present in cluster 1, is expected to have a wide variety of trackers embedded and thus, it is clearly far from other topics.

**Findings:** Overall, the two sides of clustering revealed different views. Users, due to content they are presented, can perceive the majority (75%) of topical subpages in the same way. Many of them have similar textual content and are clustered together, with the exceptions of *Art, Culture, and Society, International, Entertainment* and *Multimedia*, with the last two being clustered together due to very similar content. On the contrary, embedded TP trackers, tend

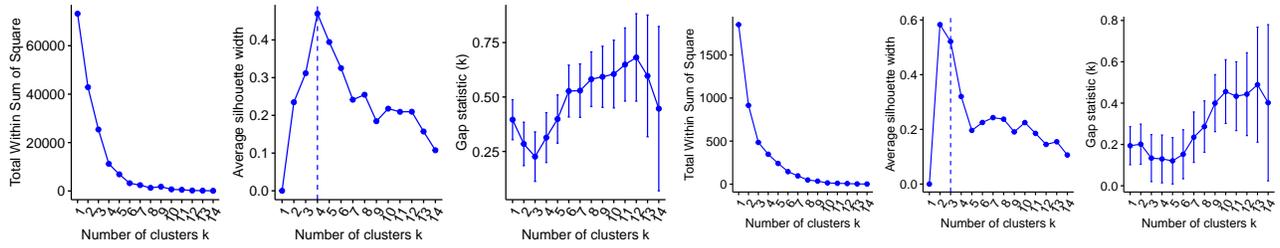
to track specific subpages simultaneously, revealing possible user profiling in action (for e.g., users interested in *Local, Regional* and *Politics*).

## 5 CONCLUSION AND FUTURE WORK

Our study is the first step in trying to understand the “Wild East” of web tracking in the Indian context. We aimed to cover two gaps in the tracking ecosystem. First, operationalizing the identification of topics using URLs of topical subpages. Second, understanding tracking on the identified topical subpages of the Indian news websites. In general, we find that the characteristics of tracking are different on topical subpages when compared to Homepages of the same news websites. We summarize our findings as follows:

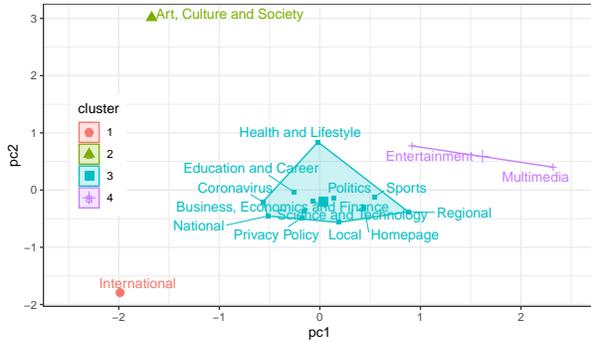
- Our proposed DiBETS model performs well for identification of news topics using only the URLs as input, with minimal user interventions.
- The tracking on topical subpages of Indian news websites are substantially different from their Homepage: more than 66% of topics have more median number of cookies than Homepages.
- In case of top ten news websites, the majority of topics have more distinct third-parties than Homepage, and in general, topics have higher Advertising and Fingerprinting cookies than Homepages.
- We found third-parties that appear to track specific subpages. There are 12% of third-parties having 100% preferential attachment towards specific topics. Further, there is topical relatedness when it comes to third-party trackers’ interest in users who view a set of topics together.
- Privacy policy subpages are mostly acting as placeholders with minimal interest of trackers. We found that eight websites do not have privacy page (also confirmed manually). All other websites have privacy policy in English, even the ones in regional languages.
- Large proportion of trackers are from the category of advertising. For some topics, there are additional trackers, which are not on Homepage. This mostly consists of trackers which indulge in user fingerprinting.
- Google’s *doubleclick.net* covers nearly 75% of the news websites in our data.
- Content shown to users appears similar for topical subpages. However, embedded third-party trackers tend to track specific subpages simultaneously, revealing possible user profiling in action (for e.g., users interested in *Local, Regional* and *Politics*), or new topics such as *Coronavirus*.

**Ethics and Limitations:** Our study does not involve any personal data from any real user. Moreover, the study helps in raising awareness and the need for appropriate laws and restrictions for tracking in the ‘East’. We plan to release all our data, code and methods for reproducibility and extensibility purposes. We tried to make our study as complete as possible. In doing so, there exist some limitations as well. The topic classification is probability-based and sensitive to threshold setting. The analysis is based on one single (best matched) subpage per topic per website. DiBETS users may be required to manually label keywords for adding them to topical dictionary. The

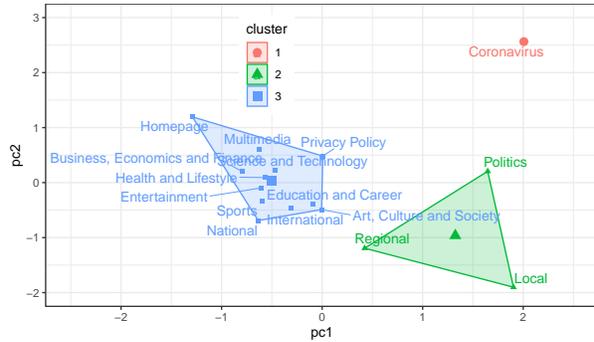


(a) Clustering metrics for K-means on PCA-reduced  $N = 2$  and different values of  $K$  on content matrix.

(b) Clustering metrics for K-means on PCA-reduced  $N = 2$  and different values of  $K$  on tracking matrix.



(c) K-Means clustering on tf-idf-based content features PCA-reduced to 2 dimensions and  $K = 4$ .



(d) K-Means clustering on third-party-based tracking features PCA-reduced to 2 dimensions and  $K = 3$ .

**Figure 6: Top: Clustering Quality Metrics: SSE (elbow method), Silhouette Coefficient, and Gap Statistic, for K-means on PCA-reduced data for shown values of  $N$  and  $K$  in (a) Content-based tf-idf features and (b) Tracking-based third-party features. Bottom: Visualization of Clustering on (c) Content-based tf-idf features and (d) Tracking-based third-party features.**

findings may be sensitive to the choice of day, time and place of crawling due to the dynamic ecosystem of tracking.

**Future Work:** We aim to expand our news websites list. Using our DiBETS model, we aim to understand topical subpage existence along party lines (Right versus Left) of news. Prior studies have shown that news leanings can be attached to the content they publish for users [10, 13]. We also aim to understand the prioritisation of topical news sharing on social media in the Indian context [1].

**ACKNOWLEDGMENTS**

N. Kourtellis has been partially supported by the European Union’s Horizon 2020 Research and Innovation Programme under grant agreements No 830927 (Concordia), No 871793 (Accordion), and No 871370 (Pimcity). These results reflect only the authors’ findings and do not represent the views of their institutes/organisations.

**REFERENCES**

[1] Pushkal Agarwal et al. 2020. Characterising user content on a multi-lingual social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 2–11.  
 [2] Pushkal Agarwal et al. 2020. Stop Tracking Me Bro! Differential Tracking Of User Demographics On Hyper-partisan Websites. In *Proceedings of the 2020 WWW conference*.  
 [3] Vibhor Agarwal et al. 2021. Under the Spotlight: Web Tracking in Indian Partisan News Websites. *arXiv preprint arXiv:2102.03656* (2021).  
 [4] Alexa. 2018. Alexa Internet. Keyword Research, Competitor Analysis, and Website Ranking. Available at <https://www.alexa.com>, accessed on 11 Jan 2021.

[5] Waqar Aqeel et al. 2020. On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement. In *Proc. of the ACM IMC*. 680–695.  
 [6] Federal Trade Commission. 1998. Children’s Online Privacy Protection Rule (COPPA). Available at <https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule>, accessed on 11 Jan 2021.  
 [7] Abhinandan S. Das et al. 2007. Google News Personalization: Scalable Online Collaborative Filtering. In *Proc. of the 16th WWW Conference* (Banff, Alberta, Canada) (*WWW ’07*). Association for Computing Machinery, New York, NY, USA, 271–280. <https://doi.org/10.1145/1242572.1242610>  
 [8] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proc. of ACM CCS 2016*.  
 [9] Steven Englehardt and Arvind Narayanan. 2021. OpenWPM Framework. Available at <https://github.com/mozilla/OpenWPM>, accessed on 26 Feb 2021.  
 [10] Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication* 57, 1 (2007), 163–173.  
 [11] Mohammed M Fouad and A Atyah Marwa. 2018. Efficient Topic Detection System for Online Arabic News. *IJCA* 180, 12 (2018), 7–12.  
 [12] Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78, 1 (2010), 35–71.  
 [13] Martin Haselmayr, Markus Wagner, and Thomas M Meyer. 2017. Partisan bias in message selection: Media gatekeeping of party press releases. *Political communication* 34, 3 (2017), 367–384.  
 [14] Disconnect Inc. 2021. Disconnect List. Available at <https://github.com/disconnectme/disconnect-tracking-protection>, accessed on 26 Feb 2021.  
 [15] J Katyayani. 2019. Hot Topic Extraction from News Websites. In *ICCB*. Springer, 297–303.  
 [16] Sandhya Keelery. 2020. Internet usage in India. Available at <https://www.statista.com/topics/2157/internet-usage-in-india/>, accessed on 26 Feb 2021.  
 [17] Barbara Krumay and Jennifer Klar. 2020. Readability of Privacy Policies. In *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 388–399.

- [18] Andrei Kutuzov et al. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proc. of the 58th Conference on Simulation and Modelling*. Linköping University Electronic Press, 271–276.
- [19] Jiahui Liu et al. 2010. Personalized News Recommendation Based on Click Behavior. In *Proc. of the 15th International Conference on IUI (Hong Kong, China) (IUI '10)*. Association for Computing Machinery, New York, NY, USA, 31–40. <https://doi.org/10.1145/1719970.1719976>
- [20] McAfee LLC. 2021. Customer URL Ticketing System. Available at <https://trustedsources.org/>, accessed on 11 Jan 2021.
- [21] Nic Newman et al. 2015. *Reuters Institute digital news report 2015: Tracking the future of news*. Reuters Institute for the Study of Journalism.
- [22] S Ramraj et al. 2020. Topic categorization of Tamil News Articles using PreTrained Word2Vec Embeddings with Convolutional Neural Network. In *2020 CISPSSE*. IEEE, 1–4.
- [23] Carsten Reinemann et al. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism* 13, 2 (2012), 221–239. <https://doi.org/10.1177/1464884911427803> arXiv:<https://doi.org/10.1177/1464884911427803>
- [24] Everett M Rogers and James W Dearing. 1988. Agenda-Setting Research: Where Has It Been, Where Is It Going? *Annals of the International Communication Association* 11, 1 (1988), 555–594. <https://doi.org/10.1080/23808985.1988.11678708> arXiv:<https://doi.org/10.1080/23808985.1988.11678708>
- [25] Karthika Subramani, Xingzi Yuan, Omid Setayeshfar, Phani Vadrevu, Kyu Hyung Lee, and Roberto Perdisci. 2020. Measuring Abuse in Web Push Advertising. *arXiv preprint arXiv:2002.06448* (2020).
- [26] Cagri Toraman and Fazli Can. 2015. A Front-Page News-Selection Algorithm Based on Topic Modelling Using Raw Text. *J. Inf. Sci.* 41, 5 (Oct. 2015), 676–685. <https://doi.org/10.1177/0165551515589069>
- [27] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the front page: Measuring third party dynamics in the field. In *Proceedings of The Web Conference 2020*. 1275–1286.
- [28] Pelayo Vallina et al. 2019. Tales from the Porn: A Comprehensive Privacy Analysis of the Web Porn Ecosystem. In *Proc. of the IMC (Amsterdam, Netherlands) (IMC)*. 245–258.
- [29] Guixian Xu et al. 2019. Research on topic detection and tracking for online news texts. *IEEE Access* 7 (2019), 58407–58418.
- [30] Chunhui Yuan and Haitao Yang. 2019. Research on K-value selection method of K-means clustering algorithm. *J—Multidisciplinary Scientific Journal* 2, 2 (2019), 226–235.