Automated Discovery of Self-Proclaimed News Providers on Facebook

Salim Chouaki
LIX, CNRS, Inria, Ecole Polytechnique, Institut Polytechnique de Paris
Palaiseau, France

Minh-Kha Nguyen
Université Grenoble Alpes
Grenoble, France

Laura Edelson
New York University
New York, NY, USA

Tobias Lauinger
New York University
New York, NY, USA

Damon McCoy
New York University
New York, NY, USA

Oana Goga
LIX, CNRS, Inria, Ecole Polytechnique, Institut Polytechnique de Paris
Palaiseau, France

ABSTRACT

The credibility of news obtained from Facebook has become a concern due to the ease with which individuals or groups can claim to be news publishers and share news-related content. Unfortunately, the lack of transparency from Facebook regarding the list of pages claiming to be news media hinders comprehensive research in this area. This paper takes a first step towards addressing this challenge by proposing an intuitive methodology that uses the GNews API and CrowdTangle to identify self-proclaimed news providers on Facebook. Through this approach, we collected data from two different periods, revealing over 26k self-proclaimed news pages in the United States, significantly more than the known 1,553 U.S.-based sources listed by Media Bias Fact Check and News Guard. Additionally, we retrieve the posting history of discovered pages. Our analysis reveals several interesting findings. The discovered pages collectively exhibit higher visibility and engagement than those listed by Media Bias Fact Check and News Guard, emphasizing the importance of studying them. We also find that, on average, 300 new self-proclaimed news pages are created every four months, that 15% of the identified news pages are news aggregators, and 57% declare to be local news. Overall, our paper shows the challenges of manually compiling an extensive list of social media news providers and emphasizes the need for automated approaches like ours.

1 INTRODUCTION

Social media platforms have changed how users consume news and stay updated on current events, with nearly half of U.S. adults now turning to social media, especially Facebook, as their primary news source [50]. This reliance on Facebook for news brings both advantages and concerns. On the one hand, it enables effortless news dissemination, democratizes access to information, and allows users to exchange ideas and opinions with people. On the other hand, many organizations have raised concerns about the platform facilitating exposure to misinformation [2, 32]. One key enabling mechanism is the ease with which anyone can claim to be a news provider and share news-related content without verification. For instance, recent reports showed the emergence of organizations aiming to influence voters during elections by claiming to be local news providers [4].

Fostering a healthy news environment requires constant monitoring and auditing of content shared by both known and less-known self-proclaimed news providers. *Unfortunately, having a comprehensive view remains impossible, as Facebook does not disclose the list of self-proclaimed news providers on the platform.* In an attempt to audit the (mostly U.S.) news media ecosystem, known journalistic agencies, Media Bias Fact Check and News Guard, have aggregated a list of 4,323 news media Facebook pages [41, 42]. As they are the only sources, many recent news-related studies have only considered established news providers listed by journalists [23, 35, 39, 49, 51]. However, we do not know to which extent these lists are comprehensive and, hence, to which extent relying studies provide an extensive view of the entire Facebook news ecosystem. Even worse, to our knowledge, there are no such lists in countries other than the U.S., which hampers both journalistic and academic auditing of the news ecosystem across countries.

In this work, *we propose a simple yet effective methodology to discover self-proclaimed news providers on Facebook* (Section 3.1). Our approach relies on the assumption that Facebook pages claiming to be (and wanting to look like) news sources typically post news-related content. Therefore, our key idea is to perform a daily crawl that: (1) exploits the GNews API [31] to get a sample of news articles published by established news media in the past 24 hours and extract a set of corresponding *keywords*; (2) uses CrowdTangle [19], an API provided by Meta, to search for Facebook posts mentioning these keywords in the past 24 hours; and (3) filters only Facebook pages that self-identify as news media. We focus in this paper on discovering news provides primarily based in the U.S. to compare the effectiveness of our method against known lists from Media Bias Fact Check and News Guard; however, our method is adaptable to identifying pages based in any country.

We deployed our methodology to gather snapshots at two periods. First, we conducted a daily detection over June 2022, identifying 19,590 Facebook news pages. Then, we conducted a retrospective detection of news pages active in October 2020 (during the U.S. presidential election) that were not deleted since, discovering 23,992 pages, including pages that have since stopped posting. Overall, our data collection enabled the discovery of *26k+ self-proclaimed news providers on Facebook based in the U.S.* Note that we simply identify pages that *claim* to be news providers without judging whether they are legitimate according to journalistic standards. This is unnecessary from an auditing perspective, as a page does not need to be legitimate to influence public opinion.

We performed several tests to evaluate the effectiveness of our method (Section 4). In terms of coverage, we find that the results

of our two deployments *cover over 95% of the pages listed by Media Bias Fact Check and News Guard*; hence, our method can catch well-known news providers. In addition, our method catches ten times more U.S.-based news providers than Media Bias Fact Check and News Guard together. If we look at the rate of new discoveries per day, our method catches 8k+ pages in the first day, 2k+ pages in the second day, and further drops to only 250 new discoveries per day after the first two weeks of daily crawls; suggesting that a one month crawl is approaching a high coverage. Regarding timeliness, we find that 90% of pages are detected in less than ten active days. The detection speed is important to capture (malicious) actors that are only active on specific periods.

Finally, we analyze several characteristics of the U.S. Facebook news ecosystem we identified, including page dynamics, organizational affiliations, posting behavior, and engagement statistics (Section 5). To perform this analysis, for each Facebook page in our dataset, we collected information about *all* the posts published from July 2017 to July 2022, including per-post engagement statistics. In total, we collected information about 191,182,320 posts. Our analysis has two end-goals: (1) offer a first insight into the much larger than previously known Facebook news ecosystem, and (2) compare pages listed by Media Bias Fact Check and News Guard with non-listed pages to analyze the relevance and need for automated discovery methods. Our results show:

(1) *Visibility and engagement*: While listed pages generally have higher individual visibility metrics like follower counts and engagement scores (in median, 86k+ followers and 6k+ interactions per week for listed pages vs. 7k+ followers and 351 interactions per week for non-listed pages), the combined totals of these metrics for non-listed pages are higher (2.65 billion followers and 113 million interactions for listed pages vs. 3.51 billion followers and 113.4 million interactions per week for non-listed pages).

(2) *Organizational affiliation*: Our analysis reveals that 44% of identified pages mention a managing organization. We retrieved 3,043 organizations, with 406 owning multiple Facebook pages. We observe that, even for organizations audited by Media Bias Fact Check and News Guard, they only review a subset of managed Facebook pages.

(3) *Posting behavior*: We find that 15% of analyzed pages are *news aggregators*, 97% of which are not listed by Media Bias Fact Check and News Guard. This category is crucial to scrutinize as such pages can easily be automated to promote specific agendas by re-sharing only information aligned with their motives. Furthermore, we find that 56% of analyzed self-proclaimed news pages are focused on *local news*, with 92% of them not listed by Media Bias Fact Check and News Guard. Local news are less likely to reach the radar of journalistic auditors and are more likely to be trusted by users [38], making this an important aspect to consider.

(4) *Page dynamics*: We find that, on average, 300 new self-proclaimed news pages are created every four months in the past 15 years. Moreover, we see two prominent peaks in page creations in 2016 and 2019, potentially linked to the U.S. presidential elections. Indeed we do find evidence of Facebook news pages that only operated in the six months before/after the U.S. 2020 presidential election.[1]

[1]For example, "Louisiana Breaking News" and "American Herald".

Overall, our finding shows the challenges of manually compiling a comprehensive list of news sources, emphasizing the need for an independent system like ours to ensure rapid and broad coverage. Given the ease of claiming to be a news provider, it is essential to go beyond established media outlets and study all Facebook pages claiming to be news providers, irrespective of their reputation, popularity, or whether they create original content or are simply content farms. We are working on building and publishing with this paper the largest database of self-proclaimed news providers on Facebook across the World. We hope this extended database will be useful for both journalistic and academic research.

2 BACKGROUND

This section provides an overview of various aspects related to Facebook news pages, including their creation, verification, and association to web domains. Additionally, it introduces the Facebook News Page Index, an archive for pages predominantly sharing news-related content on Facebook.

2.1 Facebook Pages

Facebook pages provide a platform for individuals, businesses, and organizations to build and manage their presence on the platform.

Creation of Facebook pages. Creating a Facebook page is a simple process that only requires a personal Facebook account [14]. Users can initiate page creation by visiting a designated URL [26]. They are required to provide a *name*, select a *category* aligning with the page's purpose, and can add additional information to enhance the page's description, including contact details, website links, and profile and cover photos. After completing these steps, the Facebook page becomes active and available for posting *without verifying the accuracy of the provided information, enabling users to claim any category, including news media.*

Blue badge verification. Facebook pages can request a blue badge verification from Facebook to enhance their credibility. This badge is a visual indicator proving that Facebook has confirmed the page as the authentic presence of the individual, organization, or brand it represents [9]. The process involves submitting a verification request and supplying strong proof of identity such as driver's licenses, passports, national id cards, tax filing, or utility bills [10].

While the blue badge improves the credibility of a Facebook page by ensuring accountability, it is essential to note that it does not guarantee the accuracy or reliability of the page's content.

Domain verification. Facebook pages can link to external domains to claim they represent the corresponding domain. Additionally, they can provide strong proof for this claim by verifying the associated domain. This verification process involves adding a meta tag or uploading an HTML file to the website's root directory [11, 17].

Although domain verification could be valuable for news media pages to help distinguish them from fake pages pretending to represent websites, Facebook does not mandate it. Even worse, there is no visible distinction between verified and unverified domains.

Page Publishing Authorization. In response to the 2016 U.S. presidential election controversies, Facebook introduced the *Page*

Publishing Authorization in August 2018 to enhance page accountability and prevent bad actors from hiding behind fake or compromised accounts [24]. This authorization requires the concerned pages' admins to secure their personal accounts with two-factor authentication [8] and confirm their primary country location, with non-compliance resulting in posting restrictions [6, 7]. It is applied for pages with a "high potential reach" in the U.S., India, Indonesia, and the E.U. [28].

Unlike verification badges, this authorization does not confirm a page's authenticity but focuses on securing admin accounts and confirming their location. Unfortunately, Facebook does not publicly disclose which pages undergo this process, and there is no public information on what qualifies as "high potential reach."

2.2 Facebook News Page Index

The Facebook News Page Index is an initiative introduced by Facebook to identify pages primarily publishing news-related content [13]. Page admins can apply to register their pages in this index, requiring a business and a domain verification [15]. The applications undergo an internal vetting process that considers various criteria, including sharing misinformation, violations of community standards (e.g., hate speech), engaging in clickbait and engagement bait, and other factors not disclosed by Facebook [16, 25]. Pages included in the News Page Index are exempted from the ads authorization and disclaimer processes when promoting social issues or political advertisements [12, 13].

Participation in the Facebook News Page Index is voluntary. Thus, not all self-proclaimed news pages are included in this index. Regrettably, Facebook has not made the list of pages within this index publicly available and has declined our access request.

3 METHOD AND DATA COLLECTION

This section first describes our methodology to identify self-proclaimed news providers on Facebook. For simplicity, we sometimes call them *Facebook news pages*. We then describe our deployments and the different data collections we performed for this study.

3.1 Page Discovery Method

Our *key assumption* is that news providers need to publish content discussing current events to inform their audiences and maintain interactions. Although there can be specialized news websites focusing on niche topics, news organizations that can potentially influence public opinion, and which we would like to monitor and audit, need (at least at some point) to discuss current affairs.

Our approach involves three steps (Figure 1): (1) collect keywords corresponding to current events, (2) search for Facebook posts that mention these keywords, and (3) filter the resulting Facebook posts to include only those from U.S.-based pages that share content in English and claim to be news providers. Our method is designed to perform daily the following tasks:

(1) Extracting keywords corresponding to daily news. We gather popular news headlines from Google News using the *GNews* API [31]. This API provides access to top-ranked and top-ranked-by-topic news articles across eight topics: World, Nation, Business, Technology, Entertainment, Sports, Science, and Health. We extract top-ranked and top-ranked-by-topic articles across the eight categories and limit our search to articles published in English within the past 24 hours. The median number of daily news headlines retrieved for each topic is as follows: General (37), World (67), Nation (61), Business (69), Technology (71), Entertainment (67), Sports (65), Science (30), and Health (40).

Next, we employ Yake [52], a Python library for selecting the most important keywords in a text. For every title of a news article we instruct Yake to output the most relevant two tuples made of two or more keywords. For instance, Yake generates the tuples *"investigation into Trump"* and *"social network deal"* for an article in June 2022.[2] This step yields a daily list of tuples made of two or more keywords, with a median of 1,002 tuples generated each day.

(2) Collecting Facebook posts covering daily news. We employ the CrowdTangle API, a tool provided by Meta for academics, to search for content on Facebook [18, 19]. Precisely, we use the posts-search end-point that allows retrieving posts matching given parameters and search terms [21]. For each keyword tuple obtained in the previous step, we send a request to the API and limit the search window to 24 hours. We collect all the returned posts for each request, with a median of 343k posts per day. Note that CrowdTangle returns posts from tracked pages only. The API automatically tracks all pages with more than 25,000 followers and all verified profiles, in addition to pages manually added by users [20].

For each post, the API returns various attributes such as the *post's text, published time, language,* and *engagement level*, along with information about the publisher, such as the *page's name, ID, verification status, category,* and *country of the page's admin.*

(3a) Category filtering. The prior step provides a list of Facebook pages discussing current news. Many of these pages do not claim to be news providers. We consider a Facebook page to be a self-proclaimed news provider if, on its About page, it has put one of the Facebook categories in Table 1. While some categories, like "Newspaper" or "News & media websites," are clear indicators that a page claims to be a news provider, others, such as "Media" or "Show," are less specific. We opt for a broader net to ensure high coverage and avoid missing relevant pages, particularly since many news providers listed by Media Bias Fact Check or News Guard have a general "Media" category on Facebook.

(3b) Location and language filtering. We filter U.S.-based pages that share English content. Note that, this method can be adapted for pages from different locations publishing in various languages.

We first enhance the attributes describing these pages by leveraging the Facebook Ad Library, a publicly accessible platform listing Facebook ads [27]. This library provides details about advertiser pages such as the *the name and country of the organization that manages the page and the main language used in its posts.* Each page has its dedicated web page within the Ad Library site, accessible via a specific URL format.[3] Importantly, we discover that the Ad Library provides information for all pages, including those that have never promoted ads on Facebook. We verified this with a test using

[2] https://www.axios.com/2022/06/13/government-expands-investigation-trump-social-network-deal

[3] https://www.facebook.com/ads/library/?active_status=all&ad_type=all&country=ALL&view_all_page_id={page_id}

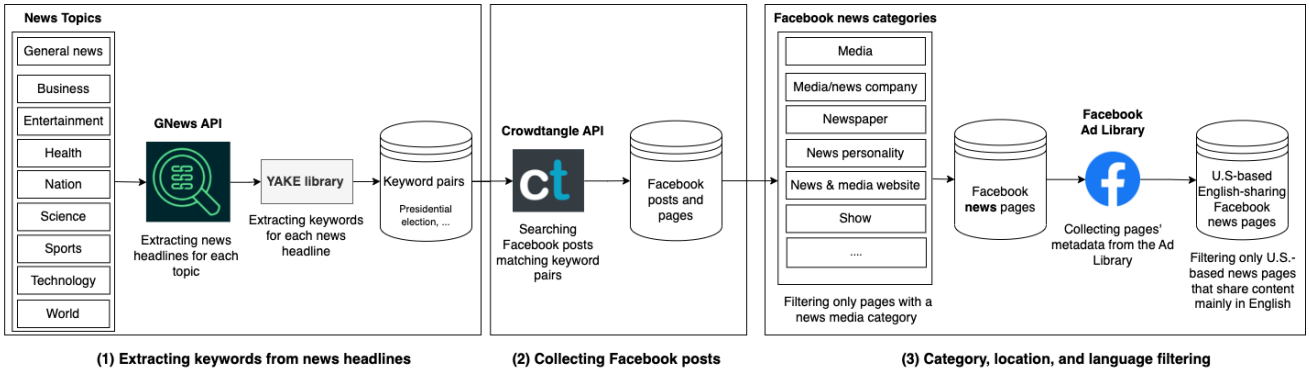

Figure 1: Diagram representing the full methodology for discovering self-proclaimed news providers' Facebook pages.

Facebook Category	News Guard & Media Bias Fact Check	Snapshot June 2022	Snapshot October 2020	Overlap
Broadcasting & media company	283	2,864	3,480	2,561
Media	11	447	614	354
Media/news company	421	6,363	7,588	5,346
Newspaper	448	2,580	3,238	2,349
Newsstand	0	8	12	6
News personality	10	2,747	3,699	2,265
News & media website	379	4,412	5,124	3,788
Show	0	129	168	89
Social Media Agency	1	40	69	24
All pages with a news category	1,553	19,590	23,992	16,782
Other categories	1059	0	0	0

Table 1: Facebook categories related to news media and the corresponding number of pages in Media Bias Fact Check and News Guard listings, SNAPSHOT_JUNE_2022, SNAPSHOT_OCTOBER_2020, and in the overlap between the two snapshots.
Top other categories include: Nonprofit Organization, Website, Publisher, Community, Political Organization, Entertainment Website, Magazine, Interest, and Public Figure.

a newly created page, confirming its presence in the Ad Library few days after its creation.

We use Selenium [47], a Python package for automating browser interactions, to retrieve information from each page's About section in the Ad Library.[4] We have created a dedicated Facebook account for this task and have implemented randomized delays of 1 to 4 seconds between each iteration to avoid bot detection.

Then, to identify U.S.-based news pages, we use two attributes: *"topAdminCountry"* provided by CrowdTangle, indicating the page's admin's country, and *"organizationCountry"* from the Facebook Ad Library, indicating the page's organization's country. We select pages where either of these attributes has *"U.S."* as a value. Finally, to identify pages primarily using English, we use the *"mainLanguage"* attribute from the Facebook Ad Library and select only pages with the value *"en."*

3.2 Datasets

We performed our first data collection in June 2022, executing the whole process once every day from June 1st to 30th, resulting in the detection of **43,436** self-proclaimed news pages. Among these, **19,590** pages are U.S.-based and primarily share content in English. We refer to this list as **SNAPSHOT_JUNE_2022**.

We conducted a second *retrospective* data collection to identify active news pages from October 1st to 30th, 2020, a sensitive period encompassing the 2020 U.S. presidential elections. This is possible

[4]For example, we access the following URL for CNN: https://www.facebook.com/ads/library/?active_status=all&ad_type=all&country=ALL&view_all_page_id=5550296508

as both the GNews and the CrowdTangle API support historical data searches within specific date ranges. We gathered data on **46,758** active news pages, with **23,992** being U.S.-based and mainly using English. We refer to this list as **Snapshot_October_2020**. Note that CrowdTangle does not return results for pages or posts that have been deleted. As a result, the dataset we obtained might represent a subset of the pages that were available in October 2020.

Across the two data collections, we compiled a total of **55,941** distinct self-proclaimed news pages, with **26,800** being U.S.-based and mainly publishing content in English.

3.3 Collection of historical posts

For each identified page, we get its posting history between July 2017 and July 2022 - i.e., all the content they have published within this timeframe. This step is not essential for discovering Facebook news pages but is important to analyze pages' posting behavior and users' engagement with their content. For this, we use the CrowdTangle dashboard's web interface to create lists of the pages for which we want to download the posting histories. Since CrowdTangle only allows downloading files with a maximum of 10,000 posts, we have manipulated the browser to automate the process and select different pages and time ranges for each download, such that we have complete post collections. Each post is characterized by the *posting time, the editing time (if the post was edited), the textual content, the type (link post, text post, image post, video post, or live video post), the post URL, the media URL, the landing URL,* and *the engagement scores of the post.* Moreover, we have additional information about the publisher with each post, such as the *number of followers at posting time.*

We collected historical data for all pages in both the Snapshot_June_2022 and Snapshot_October_2020 datasets, covering the period from July 2017 to July 2022. In total we collected **191,182,320** posts. Note that CrowdTangle does not provide posts that have been deleted or made private. Therefore, we might have gaps in the complete posting history for certain pages.

Our news discovery code and datasets are available at https://anonymous.4open.science/r/News_discovery-3B08.

4 VALIDATION

Our method aims to capture self-proclaimed news provides addressing current events in their posts. For this, we rely on external APIs and several imperfect heuristics that can impact the effectiveness of the method. Hence, this section investigates: (1) the extent to which the method can capture a comprehensive list of news media pages; (2) the speed of capturing active Facebook news pages; and (3) the extent to which the captured news pages are indeed self-proclaimed news media addressing current events.

4.1 Coverage Analysis

Ideally, we would want a method capable of identifying *all* active self-proclaimed news providers on Facebook. However, there is no existing ground truth to evaluate against. Therefore, we employ two proxies to measure the coverage of our method: (a) the extent to which it can capture well-known news media and (b) the rate at which it discovers new unseen pages. A high discovery rate indicates the difficulty in achieving comprehensive coverage since there will inevitably be more pages to discover, while a low rate suggests we may be close to achieving high coverage.

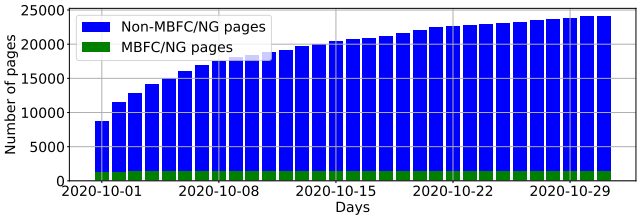

Figure 2: Cumulative number of pages detected by our method each day in October 2020. In green are pages listed by Media Bias Fact Check or News Guard, and in blue are pages not listed.

We acquired a list of well-known news providers on Facebook from Edelson et al. [23], a study aggregating news domains listed by Media Bias Fact Check (MBFC) and News Guard (NG) and their corresponding Facebook pages. This list was compiled in July 2020 and contains 4,323 news media Facebook pages. Upon verification, many of these pages are not U.S.-based (e.g., 24ur.com). To ensure a fair comparison, we excluded non-U.S.-based pages by using the *topAdminCountry* and *organisationCountry* fields, provided by CrowdTangle, to retain 2,624 U.S.-based Facebook pages. Furthermore, the MBFC/NG list includes pages that do not claim to be news providers (e.g., Public Interest Legal is categorized as "Lawyer & Law Firm," and Money and Markets as an "Investing Service"). Therefore, we further filter the MBFC/NG list to include only *1,565* pages with one of the news media categories listed in Table 1. Finally, we discarded four pages for which we could not retrieve data from CrowdTangle (due to their deletion) and eight non-English-sharing pages. As a result, we have a list of *1,553 U.S.-based English-sharing Facebook news pages* that we consider to evaluate the coverage of our method.[5]

Our analysis reveals that Snapshot_June_2022 successfully captures 89% of the U.S.-based English-sharing MBFC/NG pages, while Snapshot_October_2020 captures 94% of them. The combined scope of both snapshots includes *95%* of the MBFC/NG pages, corresponding to 1,474 out of 1,553 pages. These results prove that our method can capture well-known news media and only misses 5% of them (which we further investigate in the next section).

To provide an alternative perspective, Figure 2 presents the cumulative number per day of Facebook pages detected during October 2020. The figure shows a high discovery rate in the first seven days, with 8,781 pages on the first day and 15,053 pages within the first five days. However, the discovery rate significantly dropped afterward, with an average of 250 discovered pages per day in the last two weeks of data collection. A low rate in the second part of the crawl suggests that our method may be approaching high coverage.

[5]Note that the fact that MBFC/NG contains many non-U.S. based pages and pages not claiming to be news providers does not reflect badly on our method (or theirs). First, our method can be used to extend the list to other countries, and second, we can easily extend the collection to other Facebook categories.

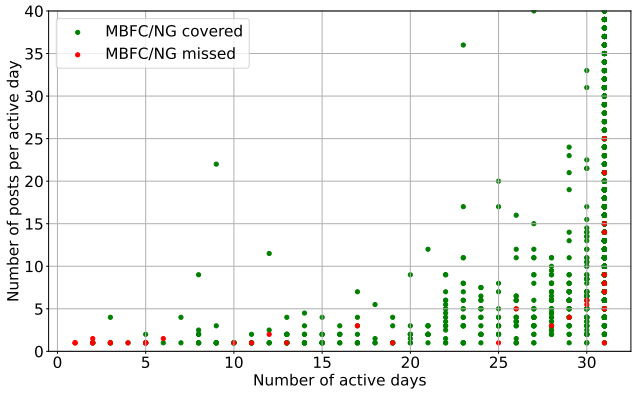

Figure 3: Number of active days and median number of posts per active day for MBFC/NG pages during October 2020.

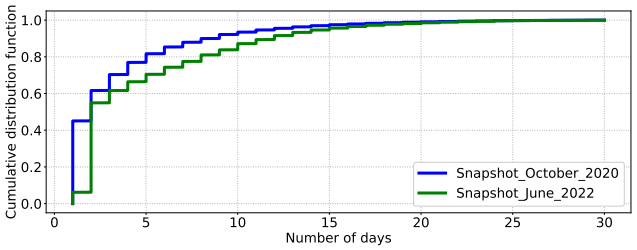

Figure 4: Cumulative distribution of the number of (active) days our method required to detect each Facebook news page in Snapshot_June_2022 and Snapshot_October_2020.

4.2 Missed Pages Analysis

The previous section shows that our method failed in detecting 79 (5%) MBFC/NG news pages. Among these, 37 pages remained inactive (i.e., did not post any content) during our two data collection periods. Hence, we are left with 42 *active* Facebook news pages we failed to detect. Our analysis of their posting activity during October 2020 reveals these pages displayed significantly lower posting frequency than the pages we successfully identified. Precisely, half of these undetected pages only published a median of one post per active day compared to 13 posts per active day for detected pages (as illustrated in Figure 3). Understandably, pages that produce limited content are less likely to meet our search filters. Thus, our ability to detect such pages is lowered.

Furthermore, we manually inspected posts from the ten most active non-detected pages (at least six posts in median per active day) to understand why they were not identified. Our findings reveal that five of these pages treat specific niche topics and did not publish content relevant to current news during our data collection. These pages include BleepingComputer, The Scientist, Community Impact, Face2Face Africa, and The Vintage News, and their respective topics, as described in their about sections, are technology, science, hyperlocal news, black history, and vintage news. Given the thematic nature of their posts, they are less likely to align with the news headline-based filtering we employ.

The remaining five pages actively posted worldwide and U.S.-related news content. However, our method did not detect them; none of their posts matched our keyword searches in CrowdTangle.

4.3 Timeliness Analysis

The dynamic nature of the Facebook news ecosystem enables malicious third parties to create several pages, share false or misleading content, and rapidly delete them. It is crucial for a method that aims to identify active news sources to detect such pages before they get deleted. Therefore, we evaluate the timeliness of our method.

To measure the time our method took to detect each page in our dataset, we count the number of active days from a page's first post in our crawling window to its detection time. We only consider days during our data collection period when pages were active, as our method cannot detect pages that do not publish anything (e.g., if a page was active only on "2020-10-01" and "2020-10-10", and we detected it on "2020-10-10", we consider that the duration for detecting this page is 2 days). Figure 4 presents the distribution of the number of (active) days our method required to detect each page within both Snapshot_June_2022 and Snapshot_October_2020. The figure demonstrates the rapid detection of most Facebook news pages, with a median detection time of two active days and more than 90% detected in less than ten active days.

4.4 Relevancy Analysis

Our method aims to identify self-proclaimed news pages sharing posts related to current events. However, it employs a few imperfect heuristics that can affect the relevancy of the pages returned:

(1) To search pages discussing current events, we rely on a list of keyword tuples. Some keyword tuples may be very general and not necessarily represent current events. For example, some extracted keywords include: arab country[6] and home sales.[7]
(2) To select self-proclaimed news providers, we refer to the categories listed in Table 1. Some of these categories are broad and can include pages not presenting themselves as news media.

To understand the relevancy of pages discovered by our method, we randomly sampled 50 pages from Snapshot_June_2022 and Snapshot_October_2020 that were not covered by MBFC/NG. We manually scrutinized the posts shared by each page in the sample to assess whether the page consistently posted content related to current news and events. We reviewed 20 random posts from each page during October 2020 or June 2022 and classified posts as news-related if they discussed current events, irrespective of the specific subject. A Facebook page was deemed a relevant news source if at least 50% of the inspected posts were news-related. Two co-authors of the paper conducted separate evaluations and the results were consistent, with *74% of the examined pages being classified as relevant news sources*. For verification, we list here the random sample of 50 pages and our relevancy classification.

One way to reduce the number of irrelevant pages is to apply stricter filters. For instance, we could consider including a Facebook page only if our method detected it on multiple distinct days, hinting

[6]Extracted from https://www.cnbc.com/2022/05/31/israel-signs-trade-deal-with-uae-its-biggest-with-any-arab-country.html
[7]Extracted from https://edition.cnn.com/2022/06/12/business/luxury-home-sales-fall-redfin/index.html

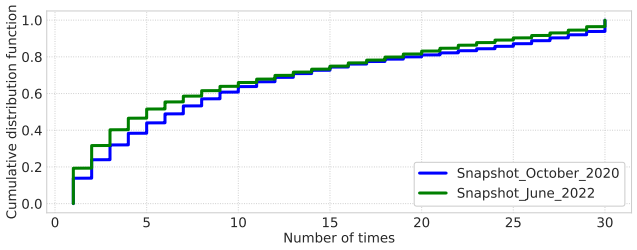

Figure 5: Cumulative distribution of the number of distinct days on which each page was detected by our method in Snapshot_June_2022 and Snapshot_October_2020.

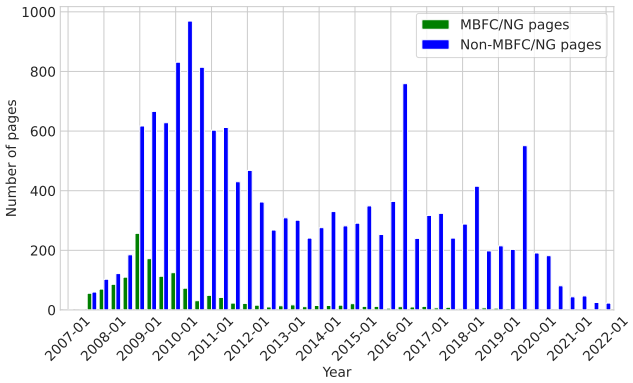

Figure 6: Creation time of Facebook pages: MBFC/NG pages vs. non-listed pages. Each point represents the number of pages created in a four-months period.

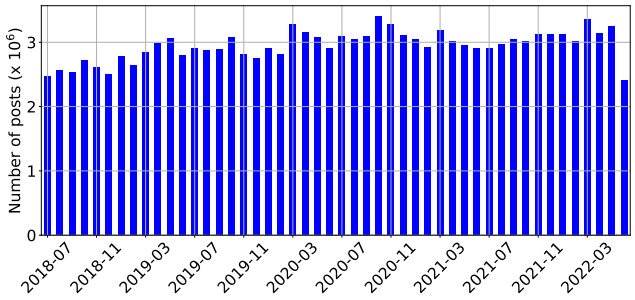

Figure 7: Combined total number of posts over all discovered pages for each month between July 2018 and July 2022

that this page shares posts about current events in a *regular* rather than *occasional* manner. Figure 5 presents the distribution of the number of distinct days on which each page was detected by our method. We can see that 19% of Snapshot_June_2022 pages and 14% of Snapshot_October_2020 pages were detected only once, and the median number of distinct days on which pages were detected is five days for Snapshot_June_2022 and seven days for Snapshot_October_2020. We performed a second manual investigation of 20 random non-listed pages selected from three distinct categories: pages detected on one or two different days, pages detected on at least five days, and pages detected on at least seven days. Our results show that 50%, 80%, and 85% pages were deemed relevant in each category. Hence, pages detected more frequently are more likely to be pertinent news sources.

Stricter filters might come at the cost of lower coverage. The list of pages detected on five or more days covers 90% of pages listed by Media Bias Fact Check and News Guard (compared to 95% when considering all pages). Our method can be adapted depending on the goal of the study.

5 NEWS ECOSYSTEM ANALYSIS

This section analyzes the characteristics of self-proclaimed news providers our method identified. We focus on non-listed pages detected in at least five daily crawls to reduce noise. For comparison, we consider the discovered listed pages without applying the same filter, as Media Bias Fact Check and News Guard have already classified them as relevant news sources. Our analysis includes 16,559 pages; 1,474 listed pages and 15,085 non-listed pages.

5.1 Dynamics

Given the risk of creating news pages to disseminate false or biased information, the first question we go after is how dynamic is the news ecosystem: (1) how many new news pages are created each year; and (2) whether they have a stable activity over time or their activity only revolves around important events such as elections.

Creation. Figure 6 presents the timeline of the creation of news pages in our dataset. The figure shows that 297 new news pages in the median were created on Facebook every four months in the past 15 years. Notably, non-listed pages tend to be more recent, with over 50% emerging after 2012, in contrast to the listed pages, where only 18% were created post-2012. We particularly see two prominent peaks in the creation time in 2016 and 2019 that might be linked to the U.S. 2016 and 2020 presidential elections.

Activity. We explore whether identified self-proclaimed news pages exhibit consistent or intermittent posting activity. Figure 7 presents a timeline of the combined number of posts across all pages per month. The figure shows that the total number of posts does not consistently increase despite the continuous creation of pages, suggesting that certain pages stop being active or are only active for specific periods. For instance, we identified 53 news pages that operated only from January 2020 to June 2021 (6 months before and after the U.S. presidential election).

These findings show the ever-changing nature of the Facebook news ecosystem, with new pages regularly emerging. Our method effectively detects these pages, particularly if used continuously.

5.2 Affiliations

To promote content associated with political and social issues on Facebook, pages are required to disclose and verify their managing organization [3]. We retrieved this information from the Facebook Ad Library for 7,277 (44%) self-proclaimed news providers pages.

We have identified 3,043 distinct organizations, of which 406 own at least two pages. Table 2 presents organizations with the largest number of Facebook pages. This table uncovers several insights.

Organization name	# Listed	# Nonlisted
Particle Media, Inc.	0	928
Planck, LLC	1	552
Gannett Satellite Information Network, LLC	88	106
Gatehouse Media LLC	72	56
Townsquare Media, INC.	3	113
Lee Enterprises Incorporated	57	50
Entercom Communications CORP.	2	81
Gray Television, INC.	54	27
BuzzFeed	2	50
Sinclair Broadcast Group Inc.	37	14
TAP Into Local LLC	1	49
Advance Local Media LLC	12	36
College Spun Media INC.	0	44
On3 Media, LLP	0	44
Insider, INC.	2	39
Alpha Media LLC	0	37
Heavy, INC.	1	35
CANTATA MEDIA LLC	1	32
Hearst	18	14
IHEARTMEDIA, INC.	0	32

Table 2: Top organizations and the number of Pages they manage that are listed by Media Bias Fact Check and News Guard, and the number of Pages not listed.

First, some news organizations possess multiple Facebook news pages, none of which are present on Media Bias Fact Check or News Guard lists. Examples include "Particle Media, Inc." and "On3 Media, LLP." Second, even organizations audited by News Guard and Media Bias Fact Check, such as "Gatehouse Media LLC" and "Sinclair Broadcast Group Inc.," have numerous Facebook pages that are not listed. For instance, while "The National Desk - TND" is included in the MBFC/NG list, "Klew News", managed by the same organization (Sinclair Broadcast Group Inc.), is not listed. These findings underscore the relevance of the self-proclaimed news sources identified by our method.

5.3 Types

The section explores various categories of self-proclaimed news providers, including news aggregators and local news sources.

News Aggregators. Recent reports have raised concerns about the rise of news aggregators, pages that republish news from various sources without creating original content, potentially driven by specific agendas and selectively sharing information aligning with their motives [5]. This section investigates the prevalence of news aggregators among self-proclaimed news providers.

We analyze landing URLs in pages' posts from July 2017 to July 2022. We first unshorten links to various URL shortening services to obtain the actual landing URL.[8] Then, we extract the distinct domains and compute the proportion of posts leading to each domain. We use the "tldextract" Python package [48] and the "Public Suffix List" [45] for this purpose. Note that we only consider the registered domain, discarding the complete domain name. For instance, if a page shares posts leading to edition.cnn.com and us.cnn.com, we consider only the unique registered domain cnn.com.

We assume that news creators predominantly share posts leading to a single domain, while news aggregators share posts with URLs spanning multiple websites, lacking a predominant domain. Therefore, we classify a page as a news aggregator if it does not have a predominant landing domain, meaning no domain accounts for 50% of the page's posts.

We find that 15% of the identified pages (2,508 pages) are news aggregators. The vast majority of these aggregators (97%) were not listed in the MBFC/NG list, like "Everything Inspirational." Additionally, we find that a median news aggregator has posted URLs from 123 distinct domains, and 1% of aggregators have shared over 1,000 domains, such as "Tully News.Info."

These results underscore another dimension of the relevance of the self-proclaimed news pages identified by our method. The method allows discovering news aggregators, most of which are not listed by Media Bias Fact Check and News Guard.

Local news We explore the geographical coverage of self-proclaimed news providers. We assume that pages primarily focused on local news at a city or state level will explicitly mention the corresponding locations in their About sections. Therefore, we analyze all pages' names and About section descriptions and classify them as local if they mention a city or a state.

For this purpose, we utilize the *locationtagger* Python library [46], which employs Named Entity Recognition techniques to extract location information, such as countries, regions/states, and cities, from input text or URLs. This library provides geographical information in three categories: cities, states, and countries. If a city or state is mentioned in the page's name or about section, we classify the page as a local news source. Similarly, if a country is mentioned, we identify the page as a source of national news content.

We extracted geographical data from 57% (9,452) of the self-proclaimed news pages. Pages lacking geographical information in their names or About sections are more likely to be national or global news sources. We find that 56% of analyzed pages (9,367) are dedicated to local news, with 52% focusing on city-level news and 4% on state-level news. Noteworthy, 92% of these local news pages are not listed by Media Bias Fact Check and News Guard.

5.4 Engagement

This section analyzes the extent to which users follow and interact with content from self-proclaimed news providers, which are valuable indicators of pages' visibility and potential impact on users.

Figure 8 presents the cumulative distribution of follower counts, and Figure 9 the cumulative distribution of average weekly interactions for discovered pages. The figures show that non-listed pages generally have significantly fewer followers (7,576 in median) and engagement scores (351 interactions per week in median) than listed

[8]The list of URL shorteners we consider: bitly.com, cutt.ly, ow.ly, rebrandly.com, short-url.at, tiny.cc, tinyurl.com, t.ly, trib.al, and usehyperlink.com.

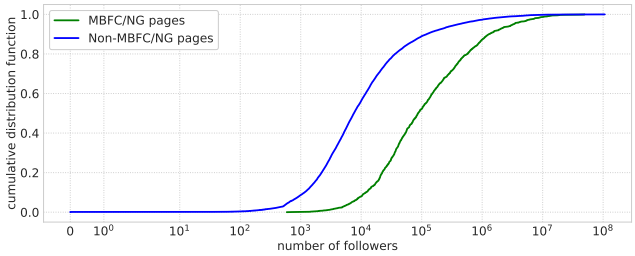

Figure 8: CDF of the number of followers for Facebook pages listed by MBFC/NG and Facebook pages not listed by MBFC/NG but discovered by our method.

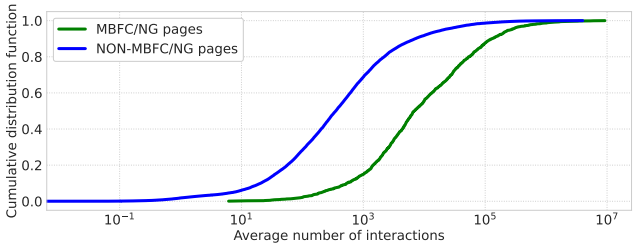

Figure 9: Cumulative distribution of the average number of interactions per active week for each Facebook news page.

pages (86,817 followers and 6,868 interactions per week). However, we find that the total followers and interaction scores across non-listed pages (3,512,253,595 followers and 113,401,864 interactions per week) are higher than those of listed pages (2,651,529,840 followers and 112,974,157 interactions per week). Hence, non-listed pages have slightly greater overall visibility (as measured by followers and engagement) than listed pages, making them important to scrutinize and consider for news and misinformation studies.

6 LIMITATIONS

Our methodology has some limitations caused by the API it relies on. First, we depend on GNews for sourcing daily news headlines and extracting search keywords. Consequently, the range of news items we can cover is tied to the news items returned by GNews. Second, to retrieve Facebook posts and pages, we rely on CrowdTangle, which exclusively returns posts from actively tracked pages. The API automatically tracks all pages with over 25,000, verified pages, and pages manually added by users. As a result, our methodology fails to identify news-related, non-verified pages with fewer than 25,000 followers that were not added by users. Finally, CrowdTangle does not provide access to posts that have been deleted or set to private. This implies that (a) we may have missed some deleted news pages in our retrospective detection, and (b) we may not have considered the complete posting history of certain pages in our analysis. Nevertheless, despite these limitations, our approach has proven to be effective in uncovering a much larger dataset of news sources on Facebook than was previously known.

7 RELATED WORK

There is a vast literature exploring online news exposure and consumption characteristics which is complementary to our work. One class of studies focused on *news shared on social media* [22, 23, 30, 33, 43]. For instance, Edelson et al. [23] employed CrowdTangle to examine the posting history of pages sourced from NewsGuard and Media Bias Fact Check and evaluate the scale and engagement scores of posts containing misinformation. On a different direction, Guess et al. [33] examined the individual-level characteristics of users associated with sharing false articles on Facebook.

A second class of studies has mainly focused on news consumption on news media websites [1, 29, 34–37, 40, 44, 49]. For instance, Horne et al. [37] present a dataset of news articles from 313 U.S. news outlets, Agarwal et al. [1] consider 103 sources to analyze the Indian news media landscape, and Scharkow et al. [49] consider 319 news domains to study the impact of news aggregators on news exposure political diversity. All these studies analyzed only a small set of news sources listed by journalists. We believe such studies could benefit from methodologies and datasets like ours, allowing for a more comprehensive news media analysis.

8 CONCLUDING DISCUSSION

Our method offers a solid starting point for future research. It provides a first step towards addressing a foundational problem for the community posed by the lack of transparency from online platforms. Facebook and similar social media do not make the index of all self-proclaimed news providers public. We did attempt to ask Facebook for such information, but our request has been denied. We hope this paper brings awareness that many news providers fly under the radar and are not covered by known lists such as Media Bias Fact Check and News Guard and pushes platforms to provide more transparency in the social media news ecosystem. Moreover, our work opens the more general question of how to define and identify news providers. For instance, pages predominately sharing news content but not self-declaring a news-related category are not considered by our method and might not be identified by journalists. We hope our extensive list will assist journalists in establishing precise criteria for what constitutes a news provider, as our dataset contains several illustrative examples.

The European Union has recently passed a legal framework, the Digital Services Act, that requires online platforms to share data with researchers and regulators for assessing systemic risks. As the European Commission is still defining the data access procedures, we believe that having an index of self-proclaimed news providers is essential to understanding and mitigating misinformation and manipulation risks and should be a high priority. Alternatively, CrowdTangle could allow searching pages based on their category. The lack of this functionality makes it challenging to identify news pages and almost impossible to have complete coverage. Moreover, we argue that all self-proclaimed news providers' pages should undergo the same verification processes as pages wanting to place political ads, providing verifiable information on the user or organization behind the page. In addition, to reduce impersonation, Facebook should require domain verification for all self-proclaimed news pages that want to list a website link in their About page.

Finally, providing aggregated audience statistics would serve as a valuable proxy for determining a page's location focus.

Overall, this paper proposes a method employing available tools for researchers, such as the GNews API and the CrowdTanble API, to identify self-proclaimed news providers on Facebook. We implement this method to discover over 26k U.S.-based news pages, significantly more than the 1,553 listed by Media Bias Fact Check and News Guard. Consequently, previous studies relying solely on these two agencies provide a restricted perspective of exposure to news, especially since we show the relevance of the additional pages we discover.

REFERENCES

[1] Vibhor Agarwal, Yash Vekaria, Pushkal Agarwal, Sangeeta Mahapatra, Shounak Set, Sakthi Balan Muthiah, Nishanth Sastry, and Nicolas Kourtellis. 2021. Under the Spotlight: Web Tracking in Indian Partisan News Websites. *Proceedings of the International AAAI Conference on Web and Social Media.*

[2] Michael Barthel, Amy Mitchell, and Jesse Holcomb. 2016. Many Americans believe fake news is sowing confusion. https://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

[3] Ben Matthews. 2022. How to get authorisation to run political and social issue ads on Facebook and Instagram. Retrieved 5 November 2023 from https://empower.agency/how-to-get-authorisation-to-run-political-and-social-issue-ads-on-facebook-and-instagram/

[4] Priyanjana Bengani. 2019. Hundreds of 'pink slime' local news outlets are distributing algorithmic stories and conservative talking points. *Columbia Journalism Review* (2019). https://www.cjr.org/tow_center_reports/hundreds-of-pink-slime-local-news-outlets-are-distributing-algorithmic-stories-conservative-talking-points.php

[5] Priyanjana Bengani. 2021. Advocacy groups and Metric Media collaborate on local 'community news'. *Columbia Journalism Review* (2021). https://www.cjr.org/tow_center_reports/community-newsmaker-metric-media-local-news.php

[6] Facebook Business. 2023. Get authorized to manage Pages with large audiences. Retrieved 5 November 2023 from https://www.facebook.com/business/m/one-sheeters/page-publishing-authorization

[7] Facebook Help Center. 2023. Get authorized to post or interact as your Page. Retrieved 5 November 2023 from https://www.facebook.com/help/1939753742723975

[8] Facebook Help Center. 2023. How wo-factor authentication works on Facebook. Retrieved 5 November 2023 from https://www.facebook.com/help/148233965247823

[9] Facebook Help Center. 2023. Request a verified badge on Facebook. Retrieved 5 November 2023 from https://www.facebook.com/help/1288173394636262

[10] Facebook Help Center. 2023. Verify Your Page or Profile. Retrieved 5 November 2023 from https://www.facebook.com/help/contact/295038365360854

[11] Meta Business Help Center. 2023. About domain verification in Meta Business Manager. Retrieved 5 November 2023 from https://www.facebook.com/business/help/286768115176155

[12] Meta Business Help Center. 2023. About News Page index. Retrieved 5 November 2023 from https://www.facebook.com/business/help/377680816096171

[13] Meta Business Help Center. 2023. About News Pages Index. Retrieved 5 November 2023 from https://www.facebook.com/business/help/377680816096171

[14] Meta Business Help Center. 2023. How to create a new Page on Facebook. Retrieved 5 November 2023 from https://www.facebook.com/business/help/1199464373557428?id=418112142508425

[15] Meta Business Help Center. 2023. Register your News Page. Retrieved 5 November 2023 from https://www.facebook.com/business/help/316333835842972

[16] Meta Business Help Center. 2023. Registration guidelines for the news Page index. Retrieved 5 November 2023 from https://www.facebook.com/business/help/270254993785210

[17] Meta Business Help Center. 2023. When to use domain verification to verify your business. Retrieved 5 November 2023 from https://www.facebook.com/business/help/245311299870862

[18] CrowdTangle. 2021. CrowdTangle access criteria. Retrieved 5 November 2023 from https://www.crowdtangle.com/request

[19] CrowdTangle. 2023. A tool from Meta to help follow, analyze, and report on what's happening across social media. Retrieved 5 November 2023 from https://www.crowdtangle.com/

[20] CrowdTangle. 2023. What data is CrowdTangle tracking? Retrieved 5 November 2023 from https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking

[21] CrowdTangle API. 2021. post-search end-point. Retrieved 5 November 2023 from https://github.com/CrowdTangle/API/wiki/Search

[22] DA Parry, BI Davidson, C Sewall JR, JT Fisher, H Mieczkowski, DS Quintana. 2021. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour* 5, 11 (2021), 1535–1547.

[23] Laura Edelson, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. 2021. Understanding Engagement with U.S. (Mis)Information News Sources on Facebook. *Proceedings of the 21st ACM Internet Measurement Conference*. https://doi.org/10.1145/3487552.3487859

[24] Facebook. 2018. Making Ads and Pages More Transparent. Retrieved 5 November 2023 from https://about.fb.com/news/2018/04/transparent-ads-and-pages/

[25] Facebook. 2019. Introducing Facebook News. Retrieved 5 November 2023 from https://about.fb.com/news/2019/10/introducing-facebook-news/

[26] Facebook. 2023. Create a Facebook page. Retrieved 5 November 2023 from https://www.facebook.com/pages/creation/

[27] Facebook. 2023. Facebook Ad Library. Retrieved 5 November 2023 from https://www.facebook.com/ads/library/

[28] Facebook Business. 2018. New Authorization for Pages. Retrieved 5 November 2023 from https://www.facebook.com/business/news/new-authorization-for-pages

[29] Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* (2016). https://doi.org/10.1093/poq/nfw006

[30] Richard Fletcher and Rasmus Kleis Nielsen. 2018. Are people incidentally exposed to news on social media? A comparative analysis. *New Media & Society* (2018). https://doi.org/10.1177/1461444817724170

[31] GNews. 2023. A Python Package that searches Google News RSS Feed and returns a usable JSON response. Retrieved 5 November 2023 from https://github.com/ranahaani/GNews

[32] Ted Van Green. 2020. Few Americans are confident in tech companies to prevent misuse of their platforms in the 2020 election. https://www.pewresearch.org/fact-tank/2020/09/09/few-americans-are-confident-in-tech-companies-to-prevent-misuse-of-their-platforms-in-the-2020-election/

[33] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* (2019). https://www.science.org/doi/abs/10.1126/sciadv.aau4586

[34] Andrew M. Guess, Pablo Barberá, Simon Munzert, and JungHwan Yang. 2021. The consequences of online partisan media. *Proceedings of the National Academy of Sciences* (2021). https://www.pnas.org/doi/abs/10.1073/pnas.2013464118

[35] Andrew M Guess, Brendan Nyhan, and Jason Reifler. 2020. Exposure to untrustworthy websites in the 2016 US election. *Nature human behaviour* (2020).

[36] Homa Hosseinmardi and Amir Ghasemian and Aaron Clauset and Markus Mobius and David M. Rothschild and Duncan J. Watts. 2021. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences* (2021). https://www.pnas.org/doi/abs/10.1073/pnas.2101967118

[37] Benjamin D. Horne, Maurício Gruppi, Kenneth Joseph, Jon Green, John P. Wihbey, and Sibel Adalı. 2022. NELA-Local: A Dataset of U.S. Local News Articles for the Study of County-Level News Ecosystems. *Proceedings of the International AAAI Conference on Web and Social Media* (2022). https://ojs.aaai.org/index.php/ICWSM/article/view/19379

[38] JOHN SANDS. 2019. Local news is more trusted than national news — but that could change. Retrieved 5 November 2023 from https://knightfoundation.org/articles/local-news-is-more-trusted-than-national-news-but-that-could-change/

[39] Ro'ee Levy. 2021. Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review* (2021). https://www.aeaweb.org/articles?id=10.1257/aer.20191777

[40] Benjamin A. Lyons, Jacob M. Montgomery, Andrew M. Guess, Brendan Nyhan, and Jason Reifler. 2021. Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences* (2021). https://www.pnas.org/doi/abs/10.1073/pnas.2019527118

[41] Media Bias Fact Check. 2023. Retrieved 5 November 2023 from https://mediabiasfactcheck.com/

[42] News Guard. 2023. Retrieved 5 November 2023 from https://www.newsguardtech.com/

[43] Anne Oeldorf-Hirsch. 2018. The Role of Engagement in Learning From Active and Incidental News Exposure on Social Media. *Mass Communication and Society* (2018). https://doi.org/10.1080/15205436.2017.1384022

[44] Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review* (2020). https://doi.org/10.37016/mr-2020-024

[45] The public suffix list. 2022. Retrieved 5 November 2023 from https://publicsuffix.org/

[46] Pypi. 2020. locationtagger. Retrieved 5 November 2023 from https://pypi.org/project/locationtagger/

[47] Pypi. 2023. Selenium 4.9.1. Retrieved 5 November 2023 from https://pypi.org/project/selenium/

[48] Pypi. 2023. The tldextract python package. Retrieved 5 November 2023 from https://pypi.org/project/tldextract/

[49] Michael Scharkow, Frank Mangold, Sebastian Stier, and Johannes Breuer. 2020. How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences* (2020). https://www.pnas.org/doi/abs/10.1073/pnas.1918279117

[50] Mason Walker and Katerina Eva Matsa. 2022. News Consumption Across Social Media in 2021. https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/

[51] Galen Weld, Maria Glenski, and Tim Althoff. 2021. Political bias and factualness in news sharing across more than 100,000 online communities. *Proceedings of the International AAAI Conference on Web and Social Media*.

[52] YAKE. 2023. Yet Another Keyword Extractor. Retrieved 5 November 2023 from https://github.com/LIAAD/yake