Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 2, March 2024, pp. 365∼378
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i2.3824

Detecting Hidden Illegal Online Gambling on .go.id Domains Using Web Scraping Algorithms

Muchlis Nurseno[1], Umar Aditiawarman[1], Haris Al Qodri Maarif[1], Teddy Mantoro[2]
[1]Universitas Nusa Putra, Sukabumi, Indonesia
[2]Universitas Sampoerna, Jakarta, Indonesia

Article Info

Article history:

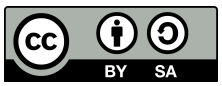
Received January 25, 2024
Revised February 20, 2024
Accepted March 01, 2024

Keywords:

Black Hat SEO
Government Website
Online Gambling
Stealthy Defacement
Web Scraper

ABSTRACT

The profitable gambling business has encouraged operators to promote online gambling using black hat SEO by targeting official sites such as government sites. Operators have used various techniques to prevent search engines from distinguishing between genuine and illegal content. This research aims to determine whether websites with the go.id domain have been compromised with hidden URLs affiliated with online gambling sites. The method used in this research is an experiment using a FOFA.info dataset containing a complete list of 450,000 .go.id domains. A web scraping algorithm developed in Python was used to identify potentially compromised websites from the targeted list by analyzing gambling-related keywords in local languages, such as 'slot,' 'judi,' 'gacor,' and 'togel'. The results showed that 958 of the 1,482 suspected.go.id sites had been compromised with an accuracy rate of 99.1%. This implies that security gaps have been exploited by illegal online gambling sites, posing a reputational risk to the government. Lastly, the scrapping algorithm tool developed in this research can detect illegal online gambling hidden in domains such as .ac.id, .or.id, .sch.id, and help authorities take necessary action.



Corresponding Author:

Umar Aditiawarman, +6281383933673
Department of Computer Science,
Universitas Nusa Putra, Sukabumi, Indonesia,
Email: umar.aditiawarman@nusaputra.ac.id

How to Cite:
M. Nurseno, U. Aditiawarman, H. Al Qodri Maarif, and T. Mantoro, "Detecting Hidden Illegal Online Gambling on .go.id Domains Using Web Scraping Algorithms", *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 365-378, Mar. 2024.

1. INTRODUCTION

Gambling activities of any form on any platform in Indonesia are prohibited and subject to Information and Electronic Transaction (ITE) law. To avoid this restriction, gambling operators employ black hat SEO techniques and exploit the website's loophole to promote online gambling activities [1, 2]. This classical method is still proven, boosts the popularity of online gambling and games, and transcends global organized crimes that already affect society and the economy [3]. As a result, online gambling has now become complex and requires comprehensive interventions to mitigate its adverse effects on society. According to the Financial Transaction Reports and Analysis Center (PPATK), the total of money circulating on online gambling in 2023 amounted to Rp 327 trillion and recorded 168 million transactions. Almost 3,3 million individuals were engaged in online gambling, and Rp 5 trillion has been transferred overseas by the operators. Between January 1, 2022, and September 6, 2023, a concerning total of 9,052 government websites were found to have been infiltrated with online gambling content, prompting how the government websites were being managed and secured. In some circumstances, these sites are hacked or defaced, and the web manager is unaware, or the pages are no longer maintained [4]. Despite global efforts to combat web defacements related to illegal activities, including online gambling, these remain elusive and challenging to eradicate [5]. This condition seriously threatens the foundational structure of government digital presence as it compromises the credibility of these service platforms and raises significant cybersecurity risks.

Although online gambling activities have compromised a lot of official websites, research focusing on the investigation of how the crime is being organized and executed is scarce. Previous studies focused on detecting negative content within websites and social media [6]. Machine learning algorithms such as SVM and Random Forest have been used to detect negative content, including pornography, fraud, and gambling materials [7, 8]. The result showed SVM is better than Random Forest in detecting negative content with an accuracy of 97%, precision of 90%, and recall of 91% from 526 websites under investigation. However, detecting negative content may not be sufficient to prevent such crimes from occurring again. Yang et al. [9] conducted a comprehensive analysis of illegal online gambling targeting Chinese users and uncovering its profit chain. They identified over 967,954 suspicious illegal gambling websites. They examined them across various dimensions, including webpage structure similarity, Search Engine Optimization (SEO) methods, abuse of Internet infrastructure, third-party online payment, and gambling groups. Min et al. [10] proposed two automatic detection systems utilizing spam SMS to identify Illegal Online Gambling (IOG) websites. Other researchers propose a hybrid multimodal data fusion-based method for identifying gambling websites by extracting and fusing visual and semantic features of the website screenshots [11]. Wang et al. [12] propose a co-training-based gambling website identification method by combining the visual and semantic features of the website screenshots. These studies essentially review the HTML structure and process the text information or URLs within the HTML for analysis. Meanwhile, Liu et al. [13] introduced a method for detecting web spam by extracting innovative feature sets from the homepage's source code and employing the random forest algorithm as the classifier. These feature sets include information extracted from the homepage's links, HTML structure, and semantic content similarity. Yang et al. [14] implemented a systematic approach to identifying and gathering infected web pages on a large scale, using keywords, search engine APIs, and a machine learning model to differentiate between infected and normal pages. Their study identified 22,939 infected pages across 2,563 domains and uncovered 8,374 new blacklisted keywords, showcasing an efficient and automated system for extensive collection and classification.

Recognizing the persistent issue of online gambling embedded on .go.id domains to bypass government restrictions and, in some cases, to enhance search engine visibility, we proposed a customed web scraping tool equipped with criteria as discussed in the previous research. Online gambling promotions in Indonesia typically leverage SEO to attain rankings on search engines. Unfortunately, they employ black hat SEO techniques by exploiting the popularity of .go.id websites. To counter black hat SEO, various detection approaches have been proposed. Web scrapers have been widespread and proven effective in solving various problems [15, 16]. Constructed using Python programming, our web scraper tool is proven to identify concealed online gambling websites within webpage content. Therefore, this study aims to determine whether websites with .go.id domains have been compromised with hidden URLs affiliated with online gambling sites. A web scrapper tool was developed to effectively detect hidden online gambling sites by examining the HTML structures. This approach will enhance the efficiency of uncovering hidden content by extracting and analyzing relevant data from websites, offering a systematic means to identify and scrutinize potential online gambling elements that might be unknowingly embedded within the web pages. The Python codes for web scraping provide a versatile solution to navigate through the complexities of various websites, aiding in the identification and analysis of concealed online gambling content.

The algorithm employed involves utilizing specific keywords to identify websites or URLs containing online gambling, which are then validated to measure their accuracy. Afterward, the compiled list of URLs is analyzed to assess the extent to which hackers have utilized .go.id domains to enhance the SEO of online gambling. Finally, this study aims to raise awareness among website administrators and government networks in Indonesia about the critical importance of consistently safeguarding their assets and remaining vigilant against covert cyber threats. Moreover, it aims to shed light on the practices of black hat SEO within the context of online gambling websites in Indonesia.

This paper is structured by first highlighting the hidden illegal online gambling phenomenon within official websites and how previous studies have encountered the issues. The gap is also identified, and the enhanced method is explained. The next section elaborates on the method used to experiment. The findings and analysis were discussed in the third section, followed by the conclusion.

2. RESEARCH METHOD

This research employs a quantitative approach based on an experimental method where web scraping algorithms were developed to detect hidden online gambling URLs from the .go.id domain. In this research, it is conceptualized that attackers engage in website defacement to inject illegal content, such as hidden illegal online gambling sites, to boost the ranking or visibility of online gambling activities. The typical implementation of black hat SEO by online gambling operators on .go.id domains is depicted in Figure 1.

Figure 1. Black Hat SEO of Online Gambling on .go.id Domain

The primary goal of this research is to develop a novel approach, namely DESLOT (Detecting Slot), which aimed at identifying potential instances of compromised websites with .go.id domains that hackers have exploited to increase the visibility and SEO of online gambling sites so that they can promote their websites on search engines. To achieve this, we leverage the results obtained from Google dork queries, which focus on websites under the .go.id domain and contain text or words similar to 'slot.' This method allows us to locate web pages that may have been defaced to facilitate illicit online gambling promotion. The web scraper data was collected between October 25 to 26, 2023. The research flow is described in Figure 2.

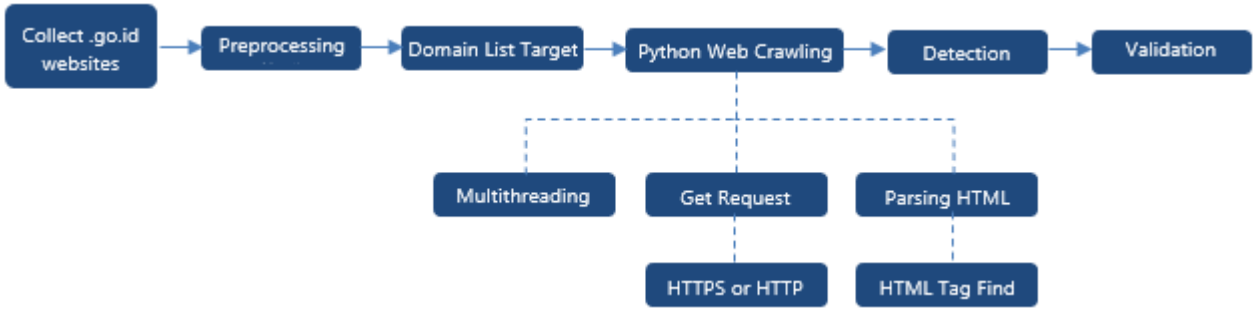


Figure 2. Research Flow

2.1. Variable Operations

As part of our investigative process, we acquire a list of domain names that end with ".go.id". This list is then used to systematically scrape the websites to determine their current status without resorting to brute force directory, subdomain crawling, or visiting links within the HTML content. Once we have determined whether each website is active or inactive, we pay close attention to the HTML tags on each page, specifically looking for keywords such as "slot," "judi," "gacor," and "togel." By carefully inspecting the HTML tags, we can identify instances where these keywords are used, which can provide valuable insights into the potential presence of gambling-related content that we can see at Figure 3. Our objective is to establish a comprehensive understanding of the current state of .go.id domains about gambling-related keywords. Our investigative process is designed to yield accurate and relevant findings, thereby contributing to our investigation's overall integrity and effectiveness.

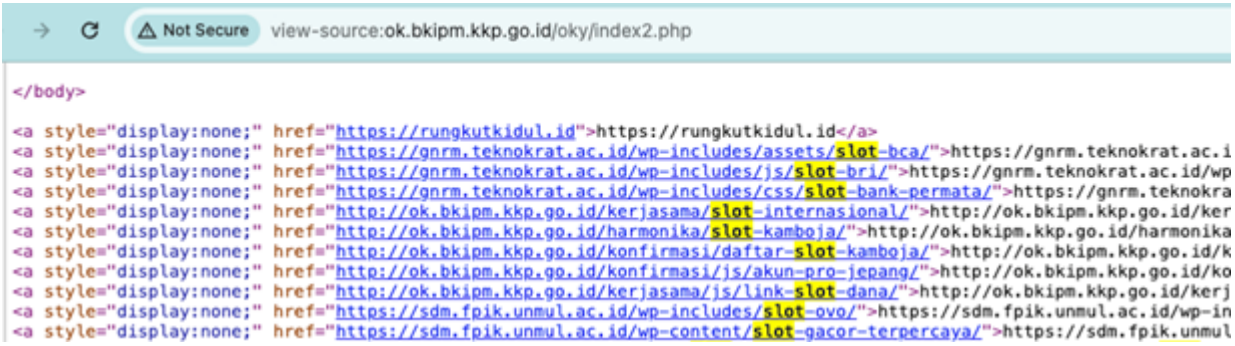


Figure 3. Affected Sites by Black Hat SEO Techniques Linked to Online Gambling Sites

2.2. The Dataset

The process of web scraping starts by identifying a target website or a group of websites from which data extraction is required. For this purpose, FOFA.info emerges as a prominent data source for web scraping endeavors. The query ”host=’.go.id’” executed on FOFA.info produces a significant number of results, approximately 450,000 records as of July 27, 2023. Subsequently, the data acquisition process was performed using the API provided by FOFA (https://en.fofa.info). The acquired data is then carefully indexed and stored in a structured format, specifically CSV. This comprehensive dataset, as shown in Figure 4, includes crucial information such as IP addresses, website titles, full domain names, main domain names, and the respective countries associated with each data entry. From 450,000 domains, data cleaning was subsequently conducted to avoid duplicates and irrelevant URLs, resulting in a list of 183,927 websites with the .go.id domain for further analysis.

Table 1. Example Result Retrieved from FOFA

Link	Domain	IP	AS_Number	AS_Organization	Country
dpk.pakpakbharatkab.go.id	pakpakbharatkab.go.id	103.114.196.246	137353	DINAS KOMINFO KABUPATEN PAKPAK BHARAT	ID
jogjaplaza.jogjaprov.go.id	jogjaprov.go.id	103.255.15.68	59151	Diskominfo DIY	ID
jbsc.jogjaprov.go.id	jogjaprov.go.id	103.255.15.93	59151	Diskominfo DIY	ID
bpttg.jogjaprov.go.id	jogjaprov.go.id	103.255.15.97	59151	Diskominfo DIY	ID
pn-batam.go.id	pn-batam.go.id	198.204.249.98	33387	NOCIX	US
mail.pa-surabaya.go.id	pa-surabaya.go.id	66.70.176.59	16276	OVH SAS	CA

The first step in the web scraping process involves consolidating the sorted full domain names into a single, comprehensive text file in .txt format. This approach ensures the accuracy and integrity of the acquired dataset, which can then be systematically scraped for further analysis. The meticulous preprocessing, efficient data organization, and systematic scraping lay the groundwork for successful data analysis, allowing valuable insights and meaningful conclusions to be extracted from the optimized dataset.

2.3. Domain Testing Techniques

A sequential scraping algorithm is developed using Python to traverse the extensive list of websites. This algorithm carefully processes each entry from the .txt file and initiates the scraping process. Starting from the first domain in the list, the algorithm systematically proceeds to the next one, ensuring thorough coverage of all websites. Important aspects of the testing process involve evaluating the connectivity and responsiveness of each website. The Python algorithm checks whether each domain is accessible and active within a 5-second time frame (Figure 5). The first attempt to access a site is made using the HTTPS protocol, and if it cannot be accessed within 5 seconds, it will be tried with the HTTP protocol for another 5 seconds. The domain is skipped if neither https nor http can be accessed within 5 seconds.

```
# Check HTTPS URL first
try:
    response = requests.get(https_url, timeout=5, verify=False, headers=headers) #tanpa allow_redirects=False
    if response.status_code == 200:
        try:
            soup = BeautifulSoup(response.content, "html.parser", from_encoding='utf-8')
        except Exception as e:
            logging.error(f"Error parsing HTML for {domain}: {e}")
            return
# If HTTPS fails, check HTTP URL
try:
    response = requests.get(http_url, timeout=5, verify=False, headers=headers) #tanpa allow_redirects=False
    if response.status_code == 200:
        try:
            soup = BeautifulSoup(response.content, "html.parser", from_encoding='utf-8')
        except Exception as e:
            logging.error(f"Error parsing HTML for {domain}: {e}")
            return
```

Figure 4. Web Connectivity Check Mechanism

In line with web scraping, the Python algorithm conducts an in-depth correlation analysis between the content of web pages and specific keywords of interest, such as 'slot', 'judi', gacor, and 'togel.' This analysis helps identify whether hackers have infiltrated the site to hide URLs or text related to online gambling. After the scraping process of all domains is completed, the Python algorithm verifies the consistency and integrity of the data across the obtained data sets and the actual web content. It performs cross-referencing with the recorded URL output and Title attributes to ensure they align with the expected values. The Python algorithm is equipped with comprehensive error-handling mechanisms to address any potential issues during the scraping process. Error logs and reports detailing any challenges encountered enable researchers to promptly identify and rectify problems, as seen in Figure 5.

```
urllib3.connectionpool - DEBUG - https://webmail.pn-tubei.go.id:2096 "GET / HTTP/1.1" 200 12324
urllib3.connectionpool - DEBUG - https://fb-ads-manager.inhukab.go.id:443 "GET / HTTP/1.1" 200 33
urllib3.connectionpool - DEBUG - https://jurung.bangka.go.id.desa.bangka.go.id:443 "GET / HTTP/1.1" 200 None
urllib3.connectionpool - DEBUG - https://erlajar.karokab.go.id:443 "GET / HTTP/1.1" 200 None
root - INFO - Processing domain: cpanel.rsuddepatihamzah.pangkalpinangkota.go.id
root - INFO - Processing domain: mail.tppkk.bandung.go.id:2095
urllib3.connectionpool - DEBUG - https://palangkaraya.go.id:443 "GET / HTTP/1.1" 200 None
urllib3.connectionpool - DEBUG - Starting new HTTPS connection (1): sipeter.bandung.go.id:2096
urllib3.connectionpool - DEBUG - Starting new HTTP connection (1): tribratanews.acehbesar.aceh.polri.go.id:80
urllib3.connectionpool - DEBUG - Starting new HTTP connection (1): webdisk.pa-makassar.go.id:80
urllib3.connectionpool - DEBUG - https://webdisk.satudata.musirawaskab.go.id:443 "GET / HTTP/1.1" 401 52
urllib3.connectionpool - DEBUG - Starting new HTTPS connection (1): webdisk.sigda.papua.go.id:443
urllib3.connectionpool - DEBUG - Starting new HTTPS connection (1): cpcalendars.disdagin.bandung.go.id:2077
root - INFO - Processing domain: webmail.pa-rantauprapat.go.id
```

Figure 5. Logging Process

In the provided Python code, the research testing techniques for web scraping are implemented to search for domains that contain gambling-related keywords ('slot', 'judi', gacor, and 'togel'). The code uses the 'requests' library for making HTTPS or HTTP requests to websites, 'BeautifulSoup' from the 'bs4' library for parsing the HTML content, 'concurrent.futures' for parallel processing, and 'csv' for writing the results to a CSV file.

In summary, The Python code demonstrates a research testing technique for web scraping to identify URLs containing gambling-related keywords, which, when the code is executed, will produce output as shown in Figure 6. It leverages libraries like 'requests' and 'BeautifulSoup' for efficient HTTP requests and HTML parsing. The use of 'concurrent.futures' ensures parallel processing of domains, improving the overall performance of the web scraping process. The results are then saved in a CSV file for further analysis and exploration.

```
  URLs: https://bontangprima.bontangkota.go.id/storage/zgacor/
  URLs: https://dpmptsp.batangharikab.go.id/sgacor/
Domain: kejari-lubuklinggau.kejaksaan.go.id - Title: kejaksaan negeri lubuklinggau
  URLs: https://kejari-manggaraibarat.kejaksaan.go.id/wp-view/slot-thailand-☕akun-pro-thaialnd-server-thailand-terbaru-2024/
  URLs: https://kejari-manggaraibarat.kejaksaan.go.id/wp-view/mpo-slot-%E2%99%9B-situs-judi-slot-gacor-online-mpo-gampang-jp-maxwin-2023/
  URLs: https://kejari-manggaraibarat.kejaksaan.go.id/wp-view/pay4d-%E2%99%9B-situs-slot-pay-4d-minimal-deposit-5000-rekomendasi-web-2023/
  URLs: https://kejari-manggaraibarat.kejaksaan.go.id/wp-view/hoki188-%E2%99%9B-link-daftar-situs-judi-gacor-gampang-jp-2023/
  URLs: https://kejari-manggaraibarat.kejaksaan.go.id/wp-view/slot303-%E2%99%9B-situs-judi-slot-gacor-303-gacor-maxin-2024/
  URLs: https://kejari-manggaraibarat.kejaksaan.go.id/wp-view/slot77-%E2%99%9B-slot-gacor-terbaik-maxin-besar-gacor-77-2024/
  URLs: https://kejari-manggaraibarat.kejaksaan.go.id/wp-view/gacor4d-%E2%99%9B-link-daftar-situs-gacor-4d-gacor-abis-2024/
Domain: perpus.bnpt.go.id - Title: perpustakaan bnpt ri
  URLs: https://revistacipa.com.br/judi-bola-euro-2024/
  URLs: https://www.dianeaskew.com/
  URLs: https://www.sistam.org/
  URLs: https://ausacademy.edu.au/slot-telkomsel/
```

Figure 6. Realtime Detection After Running Python Code

2.4. Data Analysis Strategy

To maintain data accuracy, websites subjected to web scraping undergo further validation based on their URL attributes. As shown in Table 1 below, a website is deemed Invalid if it lacks any identifiable URL or if the number of discovered URLs is less than three (3) and includes the term 'judi.' This determination is predicated on the assumption that government websites typically feature a maximum of two (2) news articles or posts related to the socialization of prevention and law enforcement campaigns against online gambling. Conversely, websites meeting these criteria are considered Valid.

Table 2. Validation Criteria

Validation	Description
Valid	There is a URL associated with online gambling content There is a title related to the term online gambling
Invalid	There is no URL or title associated with online gambling content The URL found leads to news about online gambling and its prevention. Usually, it consists of less than three urls. Apart from that, the URL is positive

The validation process plays a pivotal role in assessing the application's efficacy in identifying websites concealing URLs associated with online gambling. The validation process was also conducted to gauge its accuracy, as can be seen in Figure 7. Subsequently, the validated data undergoes meticulous scrutiny to delve into the realm of black hat SEO practices employed by online gambling websites operating under the .go.id domain.

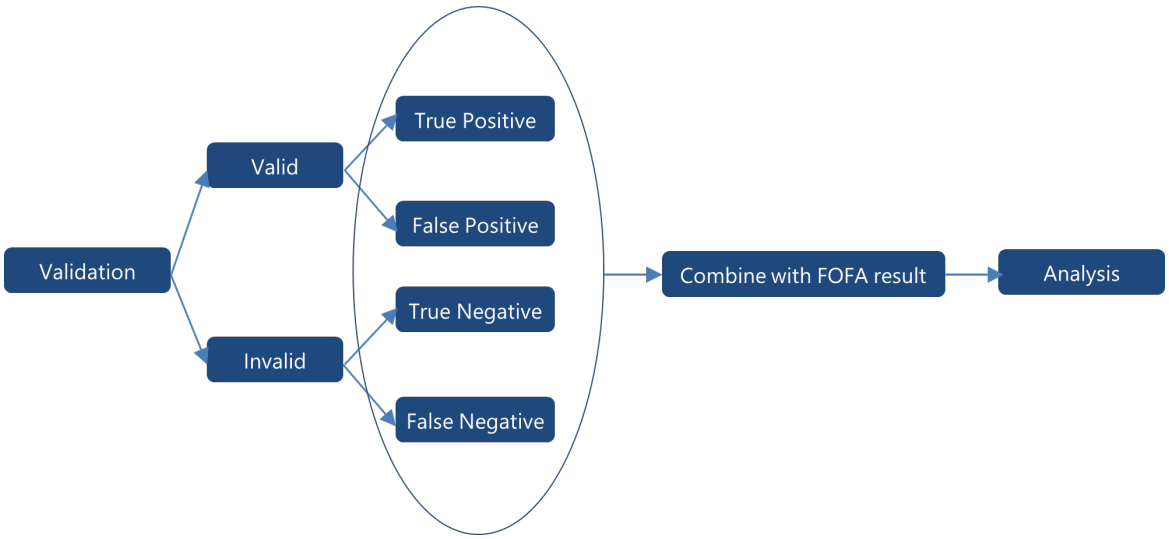


Figure 7. Validation Process

In cases where URLs are not provided, the code proceeds to validate the Title (Figure 8). It dissects the title into individual words for examination and employs regular expressions to search for specific keywords, such as 'slot', 'judi', gacor, and 'togel' while excluding the word ajudikasi. If any of these keywords are detected in the title, the code deems the website as "Valid." In situations where the URLs meet the criteria for validity, but the title does not, the domain is still considered "Valid." In determining this validation, we employ reverse logic by first identifying rows of data with invalid values and subsequently examining the valid ones.

```python
# Function to check the validity of a domain
def is_valid_domain(row):
    domain = row['Domain']
    title = row['Title']
    urls = row['URLs']

    # Checking if URLs have a non-null or non-NaN value
    if pd.notna(urls):
        urls = urls.split(';')

        # Checking if no URLs are found
        if not urls:
            return "Invalid"

        # Checking if the number of URLs is less than 3 and contains the keyword "gambling"
        if len(urls) < 3 and any("gambling" in url.lower() for url in urls):
            return "Invalid"
    else:
        # If URLs do not meet the criteria, proceed with Title validation
        if pd.notna(title):
            # Splitting the title into separate words for examination
            title_words = title.lower().split()

            # Using regular expression to find keywords in the title
            if re.search(r'\w*gambling|slot|gacor|togel\w*', ' '.join(title_words)) and not re.search(r'\bajudikasi\b', ' '.join(title_words)):
                return "Valid"

        return "Invalid"

    # If URLs are valid and Title does not meet the criteria, the domain is considered "Valid"
    return "Valid"
```

Figure 8. Validation using Python Code

As can be seen in Figure 9, for example, the invalid result in the third column (URL section) meets the following criteria: the URL is empty, the title (second column) is unrelated to online gambling, and less than three URLs are containing the keyword 'judi'. The choice of the keyword 'judi' for invalid results is because, based on observation, cybercriminals attempt to evade detection by Kominfo and website administrators by using alternative terms in their URLs, such as 'slot,' 'gacor,' and others.

Title	URLs	Validatior
polres pesisir selatan â€" polda sumbar		Invalid
pengadilan negeri lubuk pakam kelas ia - beranda		Invalid
bppd sumedang - badan promosi pariwisata daerah sumedang ðŸŽ‰	https://bppd.sumedangkab.go.id/polres-sumedang-amankan-ba	Invalid
website pengadilan negeri bangko		Invalid
polres bangkalan - tribratanews polres bangkalan	https://tribratanews.bangkalan.jatim.polri.go.id/23/10/2023/res	Invalid
pengadilan negeri ungaran - beranda		Invalid
tribrata news jawa tengah		
.: kpu kota lubuklinggau :.		
sipp		

https://tribratanews.bangkalan.jatim.polri.go.id/23/10/2023/respon-cepat-tanggapi-aduan-masyarakat-polisi-bongkar-arena-judi-sabung-ayam-di-banyuwangi/

Figure 9. Invalid Results (Uncompromised Websites)

Meanwhile, valid examples can be seen in Figure 10 below, which shows several URLs related to online gambling within the HTML body of the target website.

Title	URLs	Validatic
dinas pangan, pertanian dan perikanan kabupaten wonosobo	https://vps186137.ovh.net/tokek55/;https://	Valid
diskominfo â€" diskominfo batola	https://www.presidenslot.jdih.stiamuhamm	Valid
dinas pemadam kebakaran dan penyelamatan â€" disdamkar nat	https://kelkemenangantani.pemkomedan.gc	Valid
website	https://permainan-video.hotelsukraine.trave	Valid
dinas ket	https://slot-online.zofracobija.gob.bo/;http:,	Valid
beranda	https://rtpslot.ilovestvincent.com/products/	Valid
kecamata	https://selemadeg.tabanankab.go.id/buditoy	Valid
portal ke	https://vms.singkawangkota.go.id/slot-demc	Valid

https://kelkemenangantani.pemkomedan.go.id/d... ftar-akun-slot-gacor/;https://konijateng.id/slot777/;https://dprd.cirebonkota.go.id/wp-content/uploads/inews/;https://surfwarez.com/;https://www.hotelflora.org/;https://purworejokab.go.id/slot777.html;https://bpbd.cirebonkota.go.id/news/

Figure 10. Valid Results (Compromised Websites)

By upholding stringent validation criteria, this research endeavors to ensure the dataset's integrity. The validation process is instrumental in distinguishing between valid and invalid websites, illuminating the application's proficiency in detecting websites concealing URLs associated with online gambling. Following the completion of this accuracy assessment, the validated dataset becomes the focal point for an in-depth examination of the black hat SEO techniques employed by online gambling websites within the .go.id domain.

The valid data is then analyzed in the "URL" attribute, with the initial separation of websites under the .go.id domain. This is done to enhance the visibility of .go.id domain sites highly likely to have been compromised or utilized for black hat SEO practices related to online gambling. Subsequently, these .go.id domain websites are designated as labels to be cross-referenced with the results from FOFA.info to ascertain site identities, such as the IP addresses used. This allows us to assume that if there are .go.id domain sites with the same IP address, there is a significant likelihood that the web server of those sites has been compromised.

3. RESULT AND ANALYSIS

From the 183,927 unique websites subjected to web scraping, 1,482 were found to contain keywords such as 'slot,' 'judi,' 'gacor,' or 'togel.' However, at this stage, it cannot be definitively ascertained whether these keywords indicate infection within these sites. Following the implementation of the Python algorithm, 970 of these websites were categorized as positive or valid, of which 12 were false positives (FP) and 512 negatives, with 1 being a false negative (FN). Consequently, the count of true positives (TP) amounted to 958, and true negatives (TN) reached 511. In this case, the accuracy stands at approximately 99.1%. The accuracy of detecting websites potentially related to online gambling is calculated using Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

The seamless integration of URL and title validation in the algorithm enhances its capability to discern between valid and invalid websites accurately. Incorporating specific keywords in the 'Title' segment, indicative of potential online gambling content or activities, further augments the system's efficacy in identifying websites associated with online gambling. This advanced methodology significantly improves the overall accuracy and reliability of the algorithm's assessments, establishing it as a valuable tool for detecting concealed online gambling-related content.

Concurrently, our observations reveal a noteworthy pattern where several websites are interconnected with numerous URLs linked to online gambling. These websites strategically embed online gambling URLs using specific HTML code to bolster their visibility and search engine optimization (SEO). Employing techniques such as transparency, hidden attributes, or matching text color to the background, these websites exemplify the prevalence of black hat SEO strategies in the Indonesian online landscape, particularly focusing on utilizing .go.id domains to mask their illicit activities. These websites exhibit a clear intent to enhance the visibility and SEO of online gambling sites and unveil a complex network of interconnected domains, suggesting a coordinated effort to exploit compromised .go.id websites for this purpose.

These identified websites function as conduits to boost the ratings and SEO of websites associated with online gambling, often operating covertly under the guise of legitimate .go.id domains. Employing sophisticated HTML coding techniques to obscure online gambling-related URLs, these websites make them nearly imperceptible to casual visitors. Moreover, a significant interconnection is observed between these websites and other .go.id domains, indicating potential compromises through the creation of new paths or directories to enhance the ratings and SEO of online gambling-related websites. Notably, some of these websites even provide links to foreign online gambling sites. The statistical data related to this research can be found in Table 2.

Table 3. Summary of Processed Websites

Information	Count
Processed .go.id Websites on Web Scraper	183.927
Website Detected by Keyword	1.482
True Positive	958
URL Referring to Online Gambling	6.437
.go.id URL Referring to Online Gambling	1.514

The findings from sorting domains exposed to online gambling-related hacking and scrutinizing recorded URLs suggest that approximately 1,500 .go.id websites are implicated as tools employed by cybercriminals to boost rankings or promote online gam-

bling sites. Among these discoveries, numerous suspicious .go.id subdomains, such as slot777.sumbawabaratkab.go.id, rtp-slot.pa-sungailiat.go.id, slot-pulsa.pa-sungailiat.go.id, slot-maxwin.pa-sungailiat.go.id, slot-toto.sukabumikab.go.id, slot.sumbawabaratkab.go.id, etc., directing to the promotion of online gambling sites, further indicate the creation of new subdomains by malicious actors to evade government restrictions.

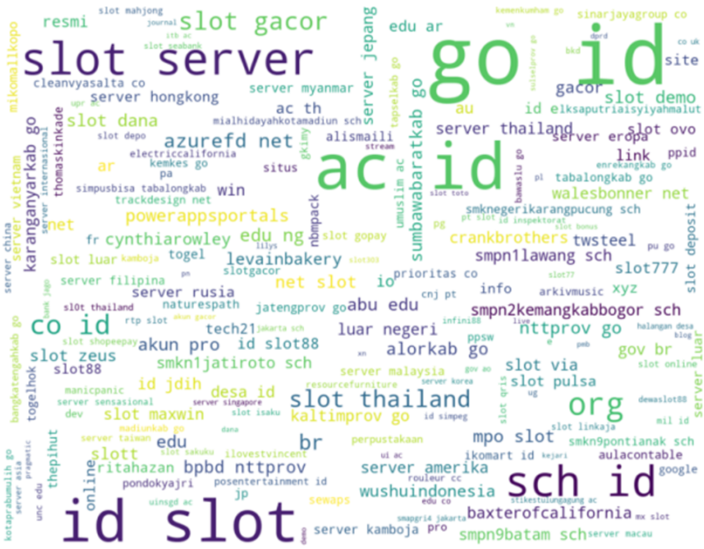


Figure 11. Wordcloud of All URLs related to Online Gambling Sites

The obtained data was subsequently processed using Python code for word cloud analysis, the results of which can be seen in Figure 11; it can be inferred that URLs related to Online Gambling predominantly include keywords such as go.id, ac.id, slot server, id slot, gacor slot, org, and others. The presence of ”ac.id” indicates that efforts to embed online gambling sites are not limited to ”.go.id” sites but also extend to sites owned by educational institutions. Additionally, various countries are mentioned, indicating them as potential locations for servers of online gambling sites abroad, such as Thailand, Russia, Hong Kong, Vietnam, Cambodia, and the United States.

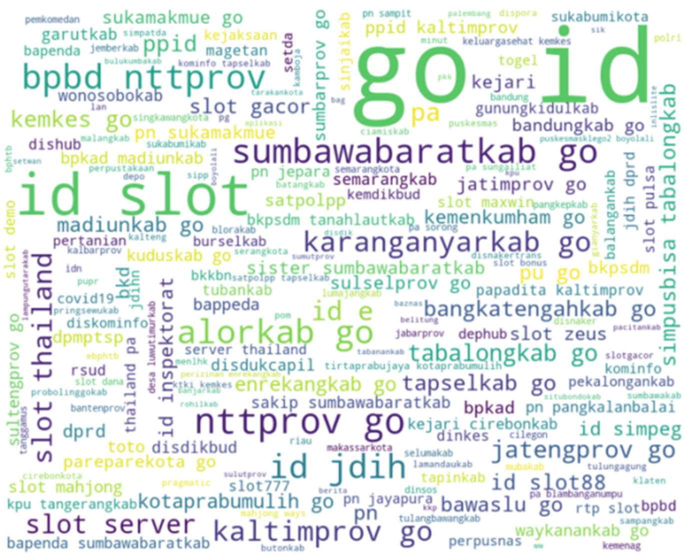


Figure 12. Wordcloud of .go.id URL Related to Online Gambling Sites

Simultaneously, the examination of the word cloud generated from ".go.id" domain URLs reveals the heightened visibility of websites associated with local government entities, as seen in Figure 12, suggesting a potential compromise of these sites. Notably, the prevalence of keywords such as "slot" within the word cloud indicates the establishment of new subdomains leveraging compromised primary domains owned by government entities. Moreover, the observed interconnectivity among these websites and other .go.id domains raises concerns, hinting at the existence of a potential network of compromised websites operating within the same domain space. The activities of this network prompt inquiries into the security and integrity of .go.id domains, necessitating further investigation into the extent of infiltration. The amalgamation of the identified 1,500 websites was subsequently cross-referenced with data obtained from FOFA.info to confirm that the websites infiltrated by illegal online gambling sites indeed constituted subdomains of a larger .go.id domain. Additionally, it was noted that several of these subdomains shared the same IP address block, implying the likelihood that the virtual machines (VM) and/or servers hosting these subdomains had been compromised.

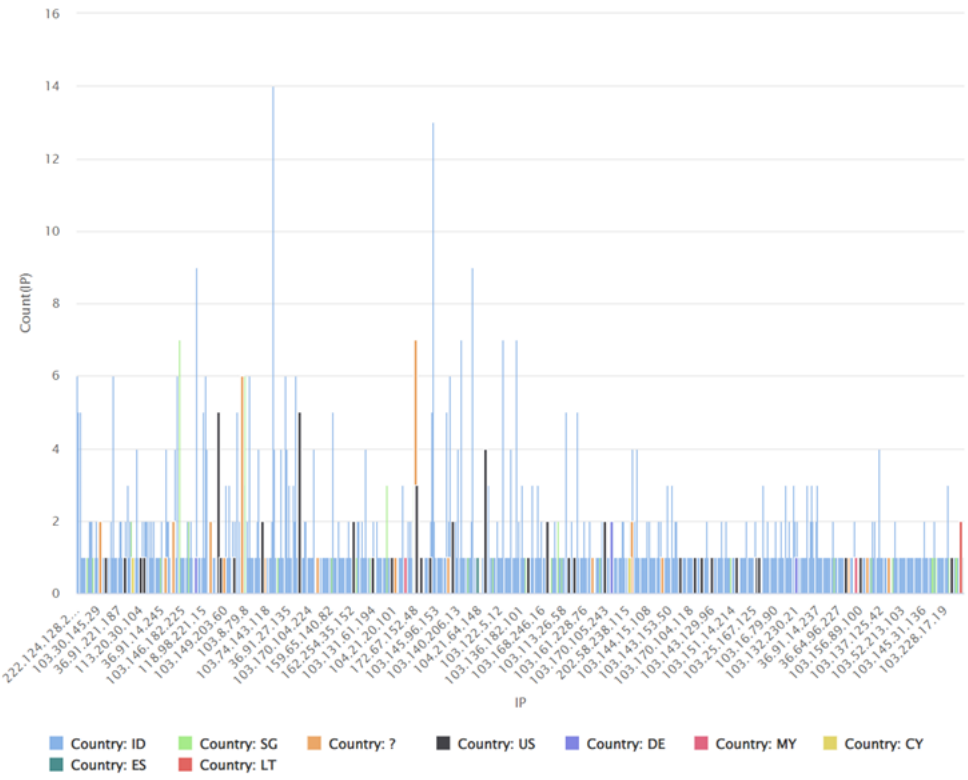


Figure 13. Wordcloud of .go.id URL Related to Online Gambling Sites

This discovery unveils a concerning pattern of illicit activity within the .go.id domain space, as the interconnection among subdomains and shared IP addresses, as seen in Figure 13, implies a network of compromised websites, raising alarms about potential breaches in the security and integrity of the .go.id domain infrastructure. This necessitates a pressing call for a comprehensive investigation to ascertain the full extent of this infiltration. Further scrutiny is imperative to comprehend the methods and motivations of cybercriminals utilizing these subdomains to boost rankings and promote illegal online gambling sites. Concurrently, efforts should be directed towards fortifying security measures and enhancing the integrity of .go.id domains to prevent future breaches and safeguard the online ecosystem from such malicious activities. In conclusion, our observations reveal a covert network operating within the .go.id domain space, discreetly elevating the ratings and SEO of online gambling-related websites through concealed URLs. Advanced HTML coding techniques underscore the prevalence of black hat SEO practices in Indonesia. The broader implications of this operation, including the interconnected compromised .go.id domains and links to foreign online gambling sites, necessitate a comprehensive investigation into the security and regulatory aspects of online activities within the country and at the international level.

Understanding the significance of these hidden links involves recognizing their role as a form of backlink manipulation. In search engine optimization (SEO), backlinks are essential elements contributing to a website's authority and visibility in search

results. A backlink is essentially a hyperlink from one website to another, acting as a vote of confidence or recommendation. When a website with a strong reputation and relevance provides a backlink to another site, it can positively impact the latter's search engine rankings.

After going through this process, we have realized that the algorithm we have developed encounters certain limitations when parsing text within HTML due to the diverse nature of text formats used by malicious actors. They utilize a variety of text types beyond the commonly used UTF-8, which necessitates additional encoding procedures. As a result, this limitation leads to the exclusion of certain URLs referring to online gambling websites. Moreover, our algorithm may still have shortcomings in the context of search engine indexing, considering that the language associated with online gambling activities is in a constant state of evolution, with new keywords emerging such as "toto," "zeus," "4D," and others. Nonetheless, the keywords we employ are considered adequate for indicating the presence of concealed activities by malicious actors in their endeavors to boost online gambling rankings.

In summary, the findings raise the need for industries, regulators, and cybersecurity experts to collaboratively address the multifaceted challenges posed by the evolving landscape of cyber threats within the Indonesian online sphere. These findings also prompt crucial recommendations to mitigate the risks associated with black hat SEO practices and bolster the security and integrity of online domains. These recommendations encompass various measures, including the development of advanced web scraping or web crawling tools capable of handling diverse text formats, fostering collaborative efforts among stakeholders, establishing clear legal frameworks, implementing continuous monitoring and evaluation mechanisms, and promoting the adoption of robust cybersecurity best practices by website owners and administrators. Additionally, conducting regular security assessments, ensuring timely software updates, enforcing strong password policies, and implementing network segmentation to shield websites against potential threats is essential. Removing unused subdomains from websites is also imperative, as neglecting supervision over inactive subdomains can create avenues for attacks. It's crucial to remember that the more doors left open, the higher the risk of vulnerabilities being exploited by malicious actors. Moreover, sensitive information about the server, such as SSH, FTP, and database credentials, should not be disclosed to unauthorized individuals or publicly. Administrators can also implement measures to address online gambling web defacement incidents by referring to documents published by the National Cyber and Encryption Agency (BSSN). Proactively safeguarding websites is key to reducing the risk of falling victim to IP address compromises and strengthening the overall cybersecurity posture. Understanding the modus operandi of these entities is essential for devising effective countermeasures to combat their activities and enhance the security of the online environment. In practice, online gambling operators swiftly create and modify websites and mechanisms, posing a significant challenge for law enforcement, particularly the Police, when dealing with cases related to online gambling [17–19].

4. CONCLUSION

With the DESLOT model, we can identify .go.id domain sites that have been infiltrated by hidden URLs for online gambling promotion. However, there are still some false positives in their execution. The strategic enhancement of our algorithm, achieved through collaborative efforts, included the integration of an additional layer of automated validation. This meticulous analysis of HTML structures containing URLs associated with online gambling played a crucial role in refining the precision of our algorithm. Employing a two-fold approach, our algorithm scrutinizes the entire HTML content and intensifies its analysis on the URL segment. The results reveal that 958 .go.id websites are compromised out of 1,482 suspected websites and provide 99.1% accuracy. After that, we delved deeper into the darker recesses of the Indonesian online landscape, revealing the widespread use of advanced HTML coding techniques and shedding light on black hat SEO practices within ".go.id" domains. The interconnected domains observed in our study highlighted a coordinated effort to exploit compromised ".go.id" websites, raising serious concerns about the security and integrity of these domains. The strategic insertion of hidden links within online gambling websites emerged as a pivotal discovery, underscoring the urgent necessity for vigilance in countering such tactics.

In addition to .go.id domains, numerous other domains warrant investigation. However, analyzing .go.id domains offers valuable insights into how black hat SEO for online gambling operates in Indonesia without detection. For example, domains such as .ac.id, .or.id, and .sch.id, which are suspected of having been extensively infiltrated by hackers to promote online gambling, also require comprehensive scrutiny. Further research in this field can shed light on the evolving tactics and techniques employed by cybercriminals in the context of online gambling and serve as a basis for developing more robust web scraping or crawling algorithms capable of handling diverse text formats.

5. ACKNOWLEDGEMENTS

The authors would like to thank everyone who contributed to this work until it was published, especially the anonymous reviewers, the chief editors, and the Matrik: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer.

6. DECLARATIONS

AUTHOR CONTIBUTION

The first and second authors are responsible for the simulations and write-up, the third author advised the framework and method used in the study, and the fourth and fifth authors are responsible for the formatting and design of the paper.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

COMPETING INTEREST

The authors have no competing financial, professional, or personal interests.

REFERENCES

[1] H. Yang, K. Du, Y. Zhang, S. Hao, H. Wang, J. Zhang, and H. Duan, "Mingling of Clear and Muddy Water: Understanding and Detecting Semantic Confusion in Blackhat SEO," in *Computer Security ESORICS 2021*. Springer, Cham, 2021, pp. 263–284, https://doi.org/10.1007/978-3-030-88418-5_13. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-88418-5_13

[2] S. Vaishy and H. Gupta, "Cybercriminals' Motivations for Targeting Government Organizations," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. Noida, India: IEEE, Sep. 2021, pp. 1–6, https://doi.org/10.1109/ICRITO51393.2021.9596104. [Online]. Available: https://ieeexplore.ieee.org/document/9596104/

[3] S. Setiawati, Pratiwi Ayu Sri Daulat, Sunarto, and SumartiniDewi, "The Urgency of Special Regulations for online Gambling in Indonesia," *International Journal of Arts and Social Science*, vol. 5, no. 7, pp. 108–115, Mar. 2023, https://doi.org/10.5281/ZENODO.7754860. [Online]. Available: https://zenodo.org/record/7754860

[4] M. Albalawi, R. Aloufi, N. Alamrani, N. Albalawi, A. Aljaedi, and A. R. Alharbi, "Website Defacement Detection and Monitoring Methods: A Review," *Electronics*, vol. 11, no. 21, p. 3573, Nov. 2022, https://doi.org/10.3390/electronics11213573. [Online]. Available: https://www.mdpi.com/2079-9292/11/21/3573

[5] R. Zhao, "The Chameleon on the Web: an Empirical Study of the Insidious Proactive Web Defacements," in *Proceedings of the ACM Web Conference 2023*. Austin TX USA: ACM, Apr. 2023, pp. 2241–2251, https://doi.org/10.1145/3543507.3583377. [Online]. Available: https://dl.acm.org/doi/10.1145/3543507.3583377

[6] A. Arora, P. Nakov, M. Hardalov, S. M. Sarwar, V. Nayak, Y. Dinkov, D. Zlatkova, K. Dent, A. Bhatawdekar, G. Bouchard, and I. Augenstein, "Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–17, Mar. 2024, https://doi.org/10.1145/3603399. [Online]. Available: https://dl.acm.org/doi/10.1145/3603399

[7] H. Syahputra and A. Wibowo, "Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 1, pp. 165–173, Mar. 2023, https://doi.org/10.26555/jiteki.v9i1.25861.

[8] Y. Chen, R. Zheng, A. Zhou, S. Liao, and L. Liu, "Automatic Detection of Pornographic and Gambling Websites Based on Visual and Textual Content Using a Decision Mechanism," *Sensors*, vol. 20, no. 14, p. 3989, Jul. 2020, https://doi.org/10.3390/s20143989. [Online]. Available: https://www.mdpi.com/1424-8220/20/14/3989

[9] H. Yang, K. Du, Y. Zhang, S. Hao, Z. Li, M. Liu, H. Wang, H. Duan, Y. Shi, X. Su, G. Liu, Z. Geng, and J. Wu, "Casino royale: a deep exploration of illegal online gambling," in *Proceedings of the 35th Annual Computer Security Applications Conference*. San Juan Puerto Rico USA: ACM, Dec. 2019, pp. 500–513, https://doi.org/10.1145/3359789.3359817. [Online]. Available: https://dl.acm.org/doi/10.1145/3359789.3359817

[10] M. Min, J. J. Lee, and K. Lee, "Detecting Illegal Online Gambling (IOG) Services in the Mobile Environment," *Security and Communication Networks*, vol. 2022, pp. 1–12, Feb. 2022, https://doi.org/10.1155/2022/3286623. [Online]. Available: https://www.hindawi.com/journals/scn/2022/3286623/

[11] C. Wang, M. Zhang, F. Shi, P. Xue, and Y. Li, "A Hybrid Multimodal Data Fusion-Based Method for Identifying Gambling Websites," *Electronics*, vol. 11, no. 16, p. 2489, Aug. 2022, https://doi.org/10.3390/electronics11162489. [Online]. Available: https://www.mdpi.com/2079-9292/11/16/2489

[12] C. Wang, P. Xue, M. Zhang, and M. Hu, "Identifying Gambling Websites with Co-training," Jul. 2022, pp. 598–603, https://doi.org/10.18293/SEKE2022-106. [Online]. Available: http://ksiresearchorg.ipage.com/seke/seke22paper/paper106.pdf

[13] J. Liu, Y. Su, S. Lv, and C. Huang, "Detecting Web Spam Based on Novel Features from Web Page Source Code," *Security and Communication Networks*, vol. 2020, pp. 1–14, Dec. 2020, https://doi.org/10.1155/2020/6662166. [Online]. Available: https://www.hindawi.com/journals/scn/2020/6662166/

[14] R. Yang, J. Liu, L. Gu, and Y. Chen, "Search & Catch: Detecting Promotion Infection in the Underground through Search Engines," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. Guangzhou, China: IEEE, Dec. 2020, pp. 1566–1571, https://doi.org/10.1109/TrustCom50675.2020.00216. [Online]. Available: https://ieeexplore.ieee.org/document/9343210/

[15] J. Schedlbauer, G. Raptis, and B. Ludwig, "Medical informatics labor market analysis using web crawling, web scraping, and text mining," *International Journal of Medical Informatics*, vol. 150, p. 104453, Jun. 2021, https://doi.org/10.1016/j.ijmedinf.2021.104453. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1386505621000794

[16] J.-C. Bricongne, B. Meunier, and S. Pouget, "Web-scraping housing prices in real-time: The Covid-19 crisis in the UK," *Journal of Housing Economics*, vol. 59, p. 101906, Mar. 2023, https://doi.org/10.1016/j.jhe.2022.101906. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S105113772200078X

[17] E. S. Hasibuan, "The Police are Indecisive: Online Gambling is Rising. Facts about the Eradication of Online Gambling in the Field," *Journal of Social Research*, vol. 2, no. 10, Aug. 2023, https://doi.org/10.55324/josr.v2i10.1405. [Online]. Available: https://ijsr.internationaljournallabs.com/index.php/ijsr/article/view/1405

[18] E. C. Listiyanto and A. Arpangi, "Implementation Effectiveness of Police Role in Eradication of Online Gaming Crime in Digital Era," *Law Development Journal*, vol. 3, no. 2, p. 362, Aug. 2021, https://doi.org/10.30659/ldj.3.2.362-370. [Online]. Available: http://jurnal.unissula.ac.id/index.php/ldj/article/view/16072

[19] M. Senjaya, "Law Enforcement of the Crime of Money Laundering That Comes from Online Gambling," *International Journal of Social Science*, vol. 2, no. 3, pp. 1641–1650, Oct. 2022, https://doi.org/10.53625/ijss.v2i3.3626. [Online]. Available: https://bajangjournal.com/index.php/IJSS/article/view/3626

[This page intentionally left blank.]