Automating Hidden Gambling Detection in Web Sites: A BeautifulSoup Implementation

Conference Paper · June 2024

DOI: 10.1109/JCSSE61278.2024.10613687

5 authors, including:

Prasert Teppap
Rajamangala University of Technology Lanna
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Panudech Tipauksorn
Rajamangala University of Technology Lanna
15 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Somnuek Surathong
Rajamangala University of Technology Lanna
5 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Prasert Luekhong
Rajamangala University of Technology Lanna
25 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Automating Hidden Gambling Detection in Web Sites: A BeautifulSoup Implementation

Prasert Teppap
Division of Electrical Engineering
Rajamangala University of Technology Lanna
Chiang Mai, Thailand
teppap@rmutl.ac.th

Panudech Tipauksorn
Division of Electrical Engineering
Rajamangala University of Technology Lanna
Chiang Mai, Thailand
panudech@rmutl.ac.th

Somnuek Surathong
Faculty of Engineering
Rajamangala University of Technology Lanna
Chiang Mai, Thailand
somnuek@rmutl.ac.th

Wirot Ponglangka
Faculty of Engineering
Rajamangala University of Technology Lanna
Chiang Rai, Thailand
wirot@rmutl.ac.th

Prasert Luekhong
College of Integrated Science and Technology
Rajamangala University of Technology Lanna
Chiang Mai, Thailand
*prasert@rmutl.ac.th

2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE) | 979-8-3503-8176-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/JCSSE61278.2024.10613687

***Abstract*—In response to escalating challenges posed by online gambling, particularly its deliberate promotion on university websites, this study enhances the capability to detect and promptly alert administrators to hidden gambling advertisements within Thai university domains. Leveraging systematic URL extraction from Google search results, we've devised a streamlined system employing BeautifulSoup for comprehensive retrieval and assessment of current gambling advertisements. Augmented with automatic reporting techniques, our system swiftly notifies network administrators of concealed gambling activity. Our primary objective is to uncover and eliminate gambling-related content within these domains. With an impressive accuracy rate of 89%, surpassing conventional methods by 53%, our system ensures instant, automatic notifications, promising to streamline the review process and bolster the cybersecurity posture of Thai universities, presenting a proactive stance against the spread of online gambling.**

***Keywords*—Hidden Gambling Detection, Web Scraping, BeautifulSoup, Automated Reporting, Cybersecurity, Thai University Websites**

I. Introduction

Particularly in the public sector, the growing volume of online gambling advertising in recent years has created significant challenges to the security and integrity of digital platforms. Government websites, usually regarded as bastions of reliability and trust, have unintentionally turned into venues for covertly placed gambling advertisements, damaging their reputation and maybe exposing users to misleading information. Such as seriously jeopardize not just the security of the internet and society as a whole, but also the integrity of government websites[1].

There are many obstacles to accurately identifying gambling advertisements on legitimate websites. Among these are sophisticated information-hiding techniques such as complex frequent URL and content changes to avoid detection[2], and the naturally complex structure of government websites that makes comprehensive scanning challenging. They not only use traditional security protocols but also automated evasion strategies, taking advantage of certain weaknesses. Bureaucratic roadblocks and inadequate maintenance funding exacerbate these problems, resulting in older systems and sluggish responses to security threats[3].

This serious issue needs a thorough approach that combines rigorous monitoring with technical progress. The volume and sophistication of modern cyber threats are making conventional methods of identifying and addressing gambling advertisements, which depend on manual inspection and reactive actions, less effective. As such, automated detection systems[4] that can rapidly detect and alert administrators to the presence of hidden gambling advertisements on government domains are essential.

Numerous researchers have extensively studied the detection of gambling websites, frequently employing machine learning techniques[5], [6] such as URLNet, Bi-LSTM, CNN-BiLSTM-Attention, NLP, ResNet34 and DRSDetector as foundational methods. These approaches have demonstrated high accuracy. However, creating datasets and training models for these purposes necessitates a substantial level of expertise. When the context or content to be detected changes, such as shifting the focus to identifying pornographic advertisements, it is essential to develop a new dataset and train a model to recognize the specific content. This preparatory work must be completed before the detection methods can be effectively utilized.

In order to bridge this gap, this study proposed to implement an automated system for recognizing and alerting Thai public sector institutions. The proposed method actively tracks website material, examines inbound links, and looks for subtle indicators of covert gambling marketing by using advanced algorithms in internet scraping and filtering. The technology will also generate daily reports for network managers, providing them with valuable information to promptly address any identified threats.

Beyond only technical advancement, this research highlights the need to uphold the integrity and dependability of digital infrastructure in the public sector. With the development and implementation of robust detection methods, this project aims to enhance defense against illicit

979-8-3503-8176-4/24/$31.00 ©2024 IEEE

online activity. This will guarantee the welfare of the people and help to preserve the credibility of the administration. The following sections of this study will cover the approaches used to build the automatic detection system, the primary algorithms used to evaluate website content, and the results of our experimental assessment. The main goal of this work is to provide exact techniques for identifying gambling advertisements, therefore reducing the need for system administrators to manually verify and guaranteeing quick automatic notifications. The study concentrated on the education industry, identifying it as the most vulnerable cyber target category by 2023[1].

There are serious concerns about public trust in the ability of government organizations to protect against online risks because gambling advertisements are so common on their websites. Our goal is to precisely identify compromised websites by conducting a thorough analysis of relevant data and approaches. Using highly developed detection algorithms, we can quickly find and report instances of illegal gambling advertising to system administrators. This proactive strategy improves the security posture of government websites by making quick remedial steps possible, strengthening public confidence in their honesty. We have divided our research groups as indicated below.

A. National Cyber Security Agency(NCSA)

The National Cyber Security Agency (NCSA) of Thailand plays a crucial role in the country's cybersecurity landscape[1]. It is responsible for protecting the national digital infrastructure and reducing cyber threats. The NCSA was created to tackle the increasing difficulties presented by cyberattacks and evolving security weaknesses. It operates under the government's supervision and serves as the main authority in charge of developing and executing thorough cybersecurity policies and plans.

Cybersecurity threats targeting educational institutions and government agencies have escalated significantly in recent years, as evidenced by the findings of the National Cyber Security Alliance (NCSA) in 2023. According to their statistics shown in Fig. 1, a total of 1,093 attacks were recorded on information services, with 632 attacks targeting education and 461 targeting government agencies. These attacks represent a substantial portion, comprising 71% of all recorded cyber incidents. The severity of these attacks underscores the critical need for effective countermeasures to mitigate their impact and ensure the uninterrupted functioning of essential services.

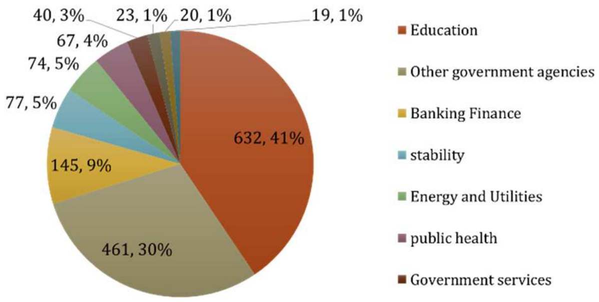


Fig. 1. Thailand's Cyber Threats: Report for 2023[1]

Moreover, the nature of cyber threats extends beyond mere infiltration, as highlighted by the prevalence of specific attack vectors observed in recent years. Notably, there has been a significant uptick in website hacks aimed at inserting online gambling advertisements, with 515 instances recorded. Additionally, there were 336 reported hacks targeting the defacement of websites, as shown in Table I. These findings underscore the urgency for comprehensive research efforts aimed at preventing website hijacking and mitigating associated risks, particularly concerning unauthorized alterations and the insertion of potentially harmful content such as gambling advertisements.

TABLE I. TOP 10 MONTHLY BREAKDOWN OF CYBER THREATS

no.	Threat patterns in 2023	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	Sum
1	Hacked Website (Gambling)	78	40	73	31	111	42	8	2	69	61	0	0	515
2	Hacked Website (Defacement)	83	60	79	17	13	20	8	5	19	32	0	0	336
3	Fake Website	20	28	50	49	26	26	13	50	13	26	0	0	301
4	WeaknessesVulnerabilities	1	1	4	6	4	10	45	6	1	150	0	0	228
5	Finance Scam	0	10	7	12	19	17	15	24	5	3	0	0	112
6	Data Leak	0	0	2	31	3	0	4	60	0	0	0	0	100
7	Data Breach	3	2	6	0	0	3	6	7	14	6	0	0	47
8	Hacked Website (Phishing)	3	6	7	1	5	4	0	4	5	3	0	0	38
9	DDoS	0	0	1	0	0	3	20	9	0	0	0	0	33
10	Ransomware	1	6	2	3	1	4	1	5	1	3	0	0	27

B. Firewall and Web Application Firewall

Firewalls(FW) and web application firewalls(WAF) constitute indispensable components of contemporary cybersecurity architectures, playing pivotal roles in safeguarding network infrastructure and web applications from unauthorized access, malicious activities, and cyberattacks. Through their proactive enforcement of access control policies and vigilant monitoring of network traffic, firewalls contribute significantly to the overall security posture of organizations, enhancing resilience against cyber threats and ensuring the confidentiality, integrity, and availability of critical assets and information.

For government entities, the internet serves as a vital channel for customer interaction and data management. However, inadequately secured internet-accessible information can have catastrophic consequences, including identity theft and financial fraud. An unsecured network, defined as one lacking credential requirements for internet access, opens the door to various malicious activities such as password and internet traffic interception, compromising company data, and theft of personal information[7].

Although FW and WAF are crucial components of cybersecurity defense, they are not comprehensive solutions and cannot address all potential threats[8]. Organizations should implement a layered security approach, integrating multiple defensive measures such as endpoint protection, security awareness training, and incident response capabilities, to effectively mitigate risks and enhance resilience against evolving cyber threats.

C. Web Scraping

Web scraping[9], [10], an automated method for extracting data from websites, has become invaluable for researchers, businesses, and developers seeking information from the internet. Among web scraping tools, Beautiful Soup[11], a Python library known for its simplicity and flexibility, has gained significant traction in academic and professional circles.

In their study, Bhatt et al[10]. delve into the realm of web scraping, also known as data scraping, which is the process of extracting information from websites for various purposes such as analysis, business data collection, and real-time data retrieval. They emphasize the significance of converting unstructured data found on web pages into a structured format for effective utilization. Additionally, they highlight the necessity of obtaining permission before extracting private data. Responsible implementation of web scraping techniques facilitates the acquisition of valuable data, facilitating tasks like market research, price comparison, trend analysis, and gaining insights into consumer behavior.

Moreover, Beautiful Soup has found widespread use in academic research across diverse disciplines. Smith et al[12]. Utilized Beautiful Soup to investigate the prevalence of misinformation in online news articles. Through systematic analysis of news website content, researchers identified and quantified instances of misinformation, shedding light on challenges associated with fake news dissemination in the digital era.

D. *Detecting Gambling*

Detecting gambling behavior in online environments has emerged as a critical research area, driven by the proliferation of online gambling platforms and the associated challenges of responsible gambling and regulatory compliance. Scholars have employed various methodologies to identify and mitigate gambling-related risks. Notably, research by Smith et al[12]. and Jonas et al[13]. has utilized data mining and machine learning algorithms to analyze user behavior patterns and predict gambling activity, enabling early intervention and support for affected individuals. Additionally, studies by Kamalov et al[14]. have employed text mining and natural language processing techniques to detect gambling-related content in online communications, facilitating targeted interventions and support services. Furthermore, advancements in computer vision, exemplified by research from Wang et al[15]. and Chen et al[16]. have enabled the automatic detection of visual cues associated with gambling in multimedia content, enhancing regulatory enforcement and responsible gambling initiatives. These studies collectively contribute to our understanding of gambling behavior detection and underscore the importance of interdisciplinary approaches to address this complex issue.

E. *LINE Notify*

LINE Notify[16]-[18] is a messaging API service developed by LINE Corporation, allowing users to send notifications and messages from various applications and services to LINE Messenger. Launched in 2016, LINE Notify provides a seamless integration between third-party applications and the LINE platform, enabling users to receive timely updates, alerts, and notifications directly on their LINE Messenger account. With its straightforward API and user-friendly interface, LINE Notify has garnered widespread adoption among businesses, developers, and individuals seeking to enhance communication and streamline notification workflows. The service offers a range of features, including customizable messages, image and sticker support, and scheduling options, making it a versatile tool for real-time communication and information dissemination.

II. METHODOLOGY

A. *The normal utilization of the Google Search Engine*

We designate this methodology as the baseline model, which entails inputting the keyword "slot" alongside the target domain. An illustrative example is depicted in Fig. 2. Subsequently, a query conducted through the Google Search Engine retrieves results associated with the presentation of gambling advertisements within the specified domain. However, it is imperative to acknowledge the potential limitations in the accuracy of the data retrieval process. This may arise due to inaccessible links, and it is plausible that the content displayed in Google search results serves solely for search engine optimization (SEO) purposes[20], [21], potentially having been removed from the website. The process of validating the existence of these gambling advertisements necessitates human verification, a task characterized by a propensity for a high error rate and prolonged verification durations.

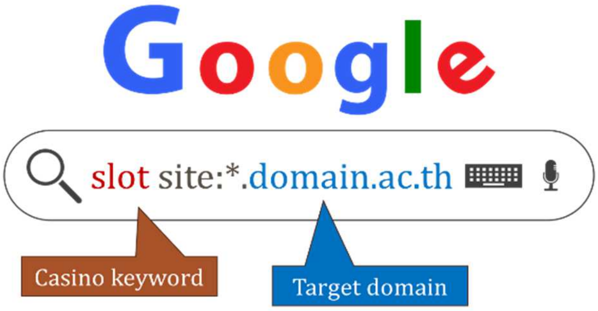


Fig. 2. Illustration of Baseline Model Implementation.

Please note that the number of search results may vary for each query. To ensure consistency in the statistical evaluation of this research, data from the top 10 search results has been exclusively included. These results are designated for subsequent statistical comparative analysis.

B. *Webpage Structure Analysis in Web Scraping*

Understanding the structure of a webpage is essential for effective web scraping, particularly when dealing with complex and dynamic websites like those associated with government domains. Webpages often comprise HTML (Hypertext Markup Language) as the foundational framework, CSS (Cascading Style Sheets) for styling, and JavaScript for interactive elements. HTML functions as the foundational framework, defining the structure and organization of information through various elements and components[22], [23].

The Document Object Model (DOM) is a crucial aspect of web scraping as it depicts the hierarchical arrangement of a webpage. It enables internet scraping tools to effectively navigate and analyze the HTML data. We utilize BeautifulSoup, a Python module, for parsing the Document Object Model (DOM). It offers a user-friendly interface for reading and altering the content of HTML and XML files, making it a very effective tool for scraping operations [24][25].

The scraping algorithm focuses largely on extracting hyperlinks (“a” tags) and relevant textual content included within tags such as “p”, “div”, and “span”. This method involves use CSS selectors and XPath to accurately locate and extract certain elements. Precise localization of certain components within the complex structure of a webpage is essential for efficient data retrieval [26].

The flowchart delineates the sequential procedures within an automated framework for detecting concealed gambling activities on websites, leveraging BeautifulSoup, a Python library tailored for web scraping and data extraction. The system orchestrates the following stages:

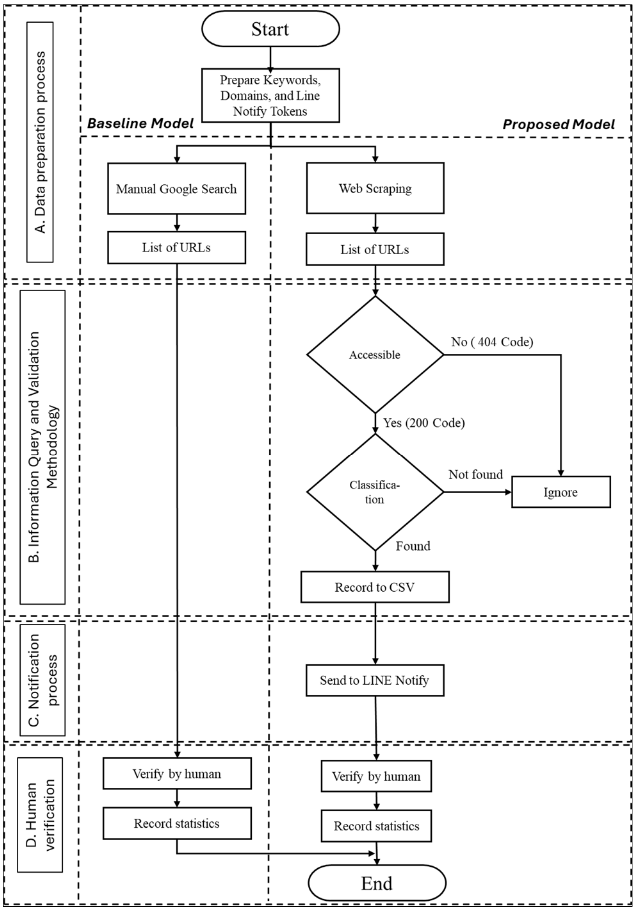


Fig. 3. Shown methodological framework.

1) Data Preparation Process

a) Collect Data Inputs: This phase encompasses gathering essential parameters requisite for the detection regimen, comprising:

Keywords: Phrases or terms indicative of gambling material, like "slot," "bonus," or "bet."

Domains: Enumerated list of website domains earmarked for scrutiny via web scraping.

LINE Notify Tokens: Credentials for LINE Notify messaging service, facilitating optional alerting.

b) Web Scaraping

Automated Web Scrutiny: This segment encompasses automated web exploration to unveil concealed gambling content on targeted websites. Its operational dynamics are as follows:

System initiates a web search, Parsing of HTML content utilizing BeautifulSoup to extricate pertinent data.

c) List of URLs: The system captures and archives a fixed number of URLs (e.g., 10) from each targeted domain, ensuring a representative sampling of webpages for subsequent analysis.

2) Information Query and Validation Methodology

a) Accessibility Validation: The system validates the accessibility status of procured URLs utilizing designated status codes:

HTTP Status Codes 200 code denotes successful retrieval of webpage content.

HTTP Status Codes 404 code indicates unavailability, potentially due to invalid URLs or webpage removal.

b) Find Keyword: In instances of accessible webpages (200 code), the system employs BeautifulSoup to comb through webpage content for predefined keywords, facilitating: Identification of gambling-related terms.Evaluation of content relevance to the detection objectives.

c) Record to CSV: Upon detection of relevant keywords, the system logs the corresponding webpage URLs into a CSV file for subsequent analysis and archival.

3) Notification Process

a) LINE Notify: Optionally, the system can dispatch notifications to designated LINE Notify channels, notifying pertinent stakeholders of potential gambling website sightings.

Fig. 3. illustrates a methodological framework designed to proficiently automate the detection of concealed gambling content prevalent across websites. Employing BeautifulSoup for web scraping control, the system embodies a potent instrument poised to swiftly unveil covertly embedded gambling-related materials on web platforms.

4) Domain expert verification

a) Verify by human : A manual validation phase is initiated, wherein human experts scrutinize the findings to discern legitimate content from false positives identified by the automated system, augmenting detection accuracy.

b) Statistical Logging: The system catalogs statistical metrics pertaining to the detection process, encompassing the tally of webpages scrutinized, gambling webpages identified, and instances of false positives.

C. Experiment

The primary aim of this project is to develop and assess an automated system tailored for detecting gambling advertisements on websites affiliated with Thai government agencies. This system leverages website scraping and data analysis tools to achieve its objectives, aiming for precise identification of covert gambling activities and prompt notification of administrators for appropriate intervention. To facilitate this study, we have specifically chosen domains associated with universities, totaling 10 institutions, as outlined in Table II.

TABLE II. SHOWN CATALOG OF SELECTED UNIVERSITIES.

Public higher education institutions	***First Selected.***	***Second Selected.***
Public universities with restricted admission	Ranking No.1	Ranking No.2
Rajamangala University of Technology	Ranking No.1	Ranking No.2
Rajabhat Universities	Ranking No.1	Ranking No.2
Autonomous university	Ranking No.1	Ranking No.2
public universities with open admissions	Ranking No.1	Ranking No.2

To provide comprehensive insights into both the target domains and the keywords utilized in this research, we meticulously curated our selection and experimentation process, which is outlined as follows.

1) University Selection: We have selectively chosen the domain of state higher education institutions from a total of 83[27], including 8 public universities with limited admissions, 9 Rajamangala University of Technology, 26 state-recognized universities, and 2 public universities without admissions restrictions. Additionally, we included 38 Rajabhat universities, selecting two from each category based on the highest number of students in each group. Consequently, the list of universities is as follows:

2) Keyword Selection: Our initial step involved examining data sourced from two reputable websites, namely wordstream.com[28] and marketkeep.com[29] , both known for aggregating popular keywords associated with casino searches. Subsequently, we conducted a comparative analysis by performing Google Search Engine queries on the top 10 keywords identified. Following this, we selected the search terms that yielded the most comprehensive information for our analytical purposes. The outcomes of this testing process are detailed in Table III below.

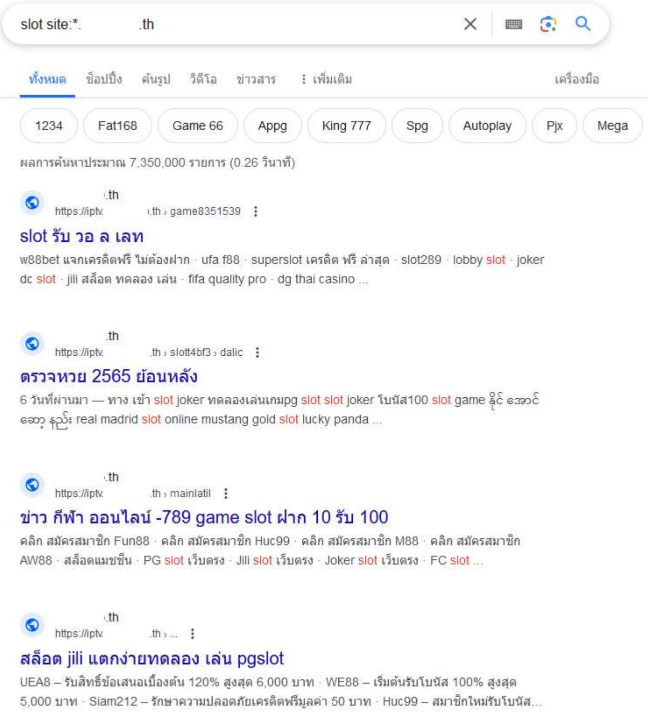
TABLE III. SHOWN KEYWORD SEARCHES

No.	*marketkeep.com*	*Searches*	*wordstream.com*	*Searches*
1	Poker	173,575	Poker	195,341
2	Casino near me	94,353	Ladbrokes	2,651
3	Jackpot city	10,486	Slots	3,119,291
4	Mohegan sun	601	Bet	2,739,808
5	Online casino	20,585	Slot machines	171,968
6	blackjack	24,431	Casino games	87,624
7	roulette	18,832	Jackpot	79,969
8	Free slots	3,036,770	Free slots	3,036,770
9	**slot**	**3,849,543**	Bonus	2,849,257
10	foxwoods	587	Poker hands	5,671

Table 3 illustrates the outcomes of 20 keyword searches sourced from both websites, which were subsequently applied across all 10 domains to aggregate the findings. This comprehensive approach aimed to identify keywords pertinent to our current research endeavor. Upon analysis of the survey data, it was observed that certain keywords yielded limited results, while the term "slot" emerged as the most frequently encountered, garnering a total of 3,849,543 views. Consequently, we elected to designate "slot" as our primary keyword for uncovering concealed gambling advertisements, particularly within websites affiliated with Thai government agencies.

The search results yielded a substantial volume of data, as illustrated in Figure 4, necessitating manual verification of each link by administrators. This process presents a significant challenge due to its labor-intensive and time-consuming nature. Our research addresses this issue by providing a solution that enables administrators to perform audits more efficiently and accurately. This approach facilitates the detection of concealed gambling advertisements within the domains under their jurisdiction.

3) System Development: Create a resilient system for extracting data from websites and filtering out unwanted content to effectively monitor website content. Utilize Beautiful Soup to efficiently identify information connected to gambling.



ผลการค้นหาประมาณ 7,350,000 รายการ (0.26 วินาที)

.th
https://iptv. .th › game8351539 ⋮
slot รับ วอ ล เลท
w88bet แจกเครดิตฟรี ไม่ต้องฝาก · ufa f88 · superslot เครดิต ฟรี ล่าสุด · slot289 · lobby slot · joker dc slot · jili สล็อต ทดลอง เล่น · fifa quality pro · dg thai casino ...

.th
https://iptv. .th › slott4bf3 › dalic ⋮
ตรวจหวย 2565 ย้อนหลัง
6 วันที่ผ่านมา — ทาง เข้า slot joker ทดลองเล่นเกมpg slot slot joker โบนัส100 slot game နိုင် အောင် ဆော့ နည်း real madrid slot online mustang gold slot lucky panda ...

.th
https://iptv. .th › mainlatil ⋮
ข่าว กีฬา ออนไลน์ -789 game slot ฝาก 10 รับ 100
คลิก สมัครสมาชิก Fun88 · คลิก สมัครสมาชิก Huc99 · คลิก สมัครสมาชิก M88 · คลิก สมัครสมาชิก AW88 · สล็อตแมชชีน · PG slot เว็บตรง · Jili slot เว็บตรง · Joker slot เว็บตรง · FC slot ...

.th
https://iptv. .th › ... ⋮
สล็อต jili แตกง่ายทดลอง เล่น pgslot
UEA8 – รับสิทธิ์ข้อเสนอเบื้องต้น 120% สูงสุด 6,000 บาท · WE88 – เริ่มต้นรับโบนัส 100% สูงสุด 5,000 บาท · Siam212 – รักษาความปลอดภัยเครดิตฟรีมูลค่า 50 บาท · Huc99 – สมาชิกใหม่รับโบนัส...

Fig. 4. Present search outcomes using the Google Search Engine.

4) Technology Deployment: Incorporate the developed technology into a tailored automated detection framework designed specifically for Thai government entities. Implement the system to thoroughly examine government agency websites to detect the existence of gambling advertisements. Set up the system to automatically generate daily reports for domain managers, providing them with practical insights to take corrective measures.

5) Evaluation and Validation: Perform comprehensive testing of the automated detection system using genuine data sourced from government agency websites. Verify the accuracy and effectiveness of the system in detecting and informing administrators about gambling advertisements. Assess the efficiency and reliability of automated systems in comparison to manual verification methods.

6) Ethical Considerations: Ensure strict adherence to ethical rules and regulations that govern the gathering and processing of data. Ensure adherence to privacy legislation to protect confidential data. Prioritize the promotion of transparency and accountability at every stage of the research process to maintain and enforce ethical standards.

Anticipated results: The expected results involve showcasing the automated system's exceptional precision in identifying gambling advertisements on Thai government websites. The technology is anticipated to expeditiously inform administrators, enabling fast corrective measures for the elimination of unauthorized content. Moreover, the study aims to make a valuable contribution to improving cybersecurity measures in the public sector and promoting responsible online behavior.

7) Confusion Matrix: The confusion matrix is a tabular representation that captures the performance of a classification algorithm[5]. It provides a visualization of the algorithm's performance by presenting the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. In this context:

True Positive (TP): The number of gambling websites correctly identified as such.

True Negative (TN): The number of normal websites correctly identified as normal.

False Positive (FP): The number of normal websites incorrectly identified as gambling.

False Negative (FN): The number of gambling websites incorrectly identified as normal.

8) Performance Metrics: Using the values from the confusion matrix, several performance metrics can be calculated:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Accuracy represents the proportion of total correct predictions (both true positives and true negatives) out of all predictions made, serving as a general measure of the classifier's performance.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Precision, also known as positive predictive value, measures the accuracy of the positive predictions. It is the proportion of true positives out of all instances predicted as positive. High precision indicates a low false positive rate.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Recall, also known as sensitivity or true positive rate, measures the ability of the classifier to identify all relevant instances. It is the proportion of true positives out of all actual positive instances. High recall indicates a low false negative rate.

$$F - measure = \frac{2x(Precision\ x\ Recall)}{Precision+Recall} \quad (4)$$

The F-measure is the harmonic mean of precision and recall, providing a single metric that balances both aspects. It is particularly useful when there is an uneven class distribution or when a balance between precision and recall is desired.

III. Result and Discussion

We conduct daily inspections by collecting statistical data and calculating results. Over a 30-day period, we identified 2,348 URLs through the Google Search Engine, averaging 81 URLs per day, or approximately 8 per domain per day. This fell short of the target of 10 URLs per domain, likely due to restrictions from the Google Search Console[30] or modifications on the websites to remove gambling advertisements. Nevertheless, the quantity of data collected was sufficient for analysis in this research. The computational methodology utilized in this analysis can be elucidated through the following experiments and equations.

A. Baseline Model

As depicted in Figure 3, we have outlined our experimental design for this research. Initially, we utilize a consistent dataset and perform searches through the Google Search Engine. We then review the search results, limiting the examination to no more than 10 URLs per domain, to determine if the URLs pertain to gambling websites.

We conduct daily checks and record the statistics. The results are as follows: True Positives (TP) = 1,242, False Positives (FP) = 253, False Negatives (FN) = 0, and True Negatives (TN) = 853, as presented in Table IV.

TABLE IV. Displays Confusion Matrix of Baseline Model.

Classification\Actual	**Gambling**	**Normal Website**
Gambling	1242	1106
Normal Website	0	0

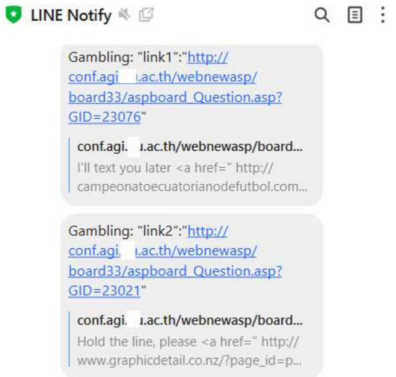
B. Proposed Model

We employed BeautifulSoup to parse the search results, utilizing the same search term ("slot") as the Baseline Model to enhance accuracy. Filtering was applied to verify the accessibility and relevance of destination links to gambling content. These URLs were classified according to the confusion matrix as follows: True Positives (TP) = 1,242, False Positives (FP) = 253, False Negatives (FN) = 0, and True Negatives (TN) = 853, as depicted in Table V.

TABLE V. Displays Confusion Matrix of Proposed Model.

Classification\Actual	**Gambling**	**Normal Website**
Gambling	1242	253
Normal Website	0	853

To facilitate prompt notification to the administrator or domain owner, we have implemented a notification system using LINE Notify. Notifications are triggered if gambling-related information is detected based on predefined search terms. The system forwards the link to the designated LINE Token. The submission format is illustrated in Figure 5.



Gambling: "link1":"http://conf.agi[redacted].ac.th/webnewasp/board33/aspboard_Question.asp?GID=23076"

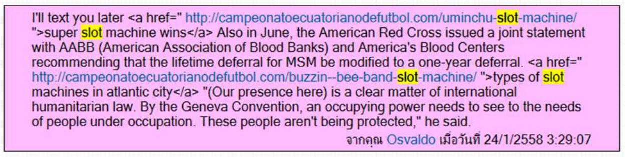
conf.agi.[redacted].ac.th/webnewasp/board...
I'll text you later <a href=" http://campeonatoecuatorianodefutbol.com...

Gambling: "link2":"http://conf.agi.[redacted].ac.th/webnewasp/board33/aspboard_Question.asp?GID=23021"

conf.agi.[redacted].ac.th/webnewasp/board...
Hold the line, please <a href=" http://www.graphicdetail.co.nz/?page_id=p...

Fig. 5. Present notification messages using LINE Notify.

Fig. 6. illustrates an additional instance from a distinct field, providing more evidence of the system effectiveness in identifying web pages connected to gambling.



I'll text you later <a href=" http://campeonatoecuatorianodefutbol.com/uminchu-slot-machine/ ">super slot machine wins</a> Also in June, the American Red Cross issued a joint statement with AABB (American Association of Blood Banks) and America's Blood Centers recommending that the lifetime deferral for MSM be modified to a one-year deferral. <a href=" http://campeonatoecuatorianodefutbol.com/buzzin--bee-band-slot-machine/ ">types of slot machines in atlantic city</a> "(Our presence here) is a clear matter of international humanitarian law. By the Geneva Convention, an occupying power needs to see to the needs of people under occupation. These people aren't being protected," he said.

จากคุณ Osvaldo เมื่อวันที่ 24/1/2558 3:29:07

Fig. 6. Another Instance of Gambling Content in an Active URL

These results were further processed into performance metrics, as detailed in Table 7.

TABLE VI. DISPLAYS THE PERFORMANCE METRICS.

Evaluation Metrics	Baseline Model	Proposed Model
accuracy	0.53	0.89
Precision	0.53	0.83
Recall	1	1
F1-Score	0.69	**0.91**

These metrics reflect the model's effectiveness in accurately detecting URLs, demonstrating a high level of precision, recall, and overall performance.

C. Discussion:

An unidentified link type within the Google Search Engine is its tendency to persistently display search results even after administrators delete associated files. This presents formidable obstacles for system administrators, particularly in scenarios involving numerous URLs, as evidenced by the error percentages depicted in Figure 5. In contrast, our research has incorporated a verification process to ascertain the presence of the destination website at designated URLs, ensuring a heightened level of accuracy. Nonetheless, occasional errors have arisen when the destination website employs the term "slot" with an alternative connotation unrelated to gambling. We elaborate on our research methodology below.

1. Exceptional Precision in Active Domains:

The software exhibited remarkable accuracy in detecting gambling-related information inside active domains. The system filtering and identification algorithms were proven successful as all URLs from these domains that underwent manual examination were found to contain gambling material.

2. Issues with Inactive URLs:

The incapacity to access some marked URLs poses a constraint in conducting a full evaluation of the system efficacy. The presence of these inaccessible URLs may suggest content removal or firewall blocking, both of which are often employed web governance measures in government domains.

3. Implications for Digital Governance:

The results emphasize the effectiveness of the program as a means for government entities and network administrators to oversee and control online gambling material. The system exceptional precision in dynamic fields implies its potential efficacy in assisting digital governance and cybersecurity initiatives.

4. Future Enhancements:

Additional enhancements to our research methodology will focus on enhancing the detection capabilities concerning unreachable or unidentified URLs resulting from content deletion or firewall restrictions. Expanding the system's capabilities to incorporate predictive analytics and machine learning techniques holds promise for augmenting accuracy and adaptability. This entails broadening the scope of keywords utilized in searches and expanding the number of domains under scrutiny to encompass a more comprehensive range of targets, a crucial consideration moving forward.

5. Broader Impact and uses:

Although primarily targeted towards government domains, the system technique may be modified for wider uses, such as surveillance in business or educational domains. The system potential influence in numerous areas involved with digital content management is emphasized by its broader application.

IV. CONCLUSION

The presence of gambling advertisements on government websites in Thailand highlights the vulnerabilities within government organizations in mitigating online risks. Our objective is to detect these advertisements on government websites and promptly notify the administrators if such content is found. We anticipate that this will lead to swift resolution of the issues, thereby restoring the efficiency of the provided services.

In summary, our research attests to an impressive accuracy rate of 89%, a substantial advancement over the 53% accuracy achieved by the "Baseline Model" method. Consequently, our study offers notable benefits by mitigating errors in reporting gambling-related links, thereby alleviating the administrative burden associated with extensive verification processes. This improvement facilitates administrators in optimizing their allocation of time and resources, enabling them to concentrate on critical tasks or conduct comprehensive data reviews with heightened precision and efficiency.

In moving forward, optimizing our research process will concentrate on increasing the number of URLs examined per domain. Enhancing the system's capabilities to incorporate predictive analytics and machine learning methods holds promise for improving accuracy and adaptability. This involves broadening the search terms and expanding the range of domains tested to ensure comprehensive coverage. These considerations are critical for future endeavors.

ACKNOWLEDGMENT

The authors would like to thank the significant educational subsidies provided by Rajamangala University of Technology Lanna during the program. Scholarships were crucial for the success of this research.

REFERENCES

[1] "Cyber threat statistics." Accessed: Mar. 17, 2024. [Online]. Available: https://www.ncsa.or.th/service-statistics.html

[2] M. Aljabri *et al.*, "Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions," *IEEE Access*, vol. 10, pp. 121395–121417, 2022, doi: 10.1109/ACCESS.2022.3222307.

[3] R. Vaarandi and S. Mases, "How to Build a SOC on a Budget," in *Proceedings of the 2022 IEEE International Conference on Cyber Security and Resilience, CSR 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 171–177. doi: 10.1109/CSR54599.2022.9850281.

[4] M. Min, J. J. Lee, and K. Lee, "Detecting Illegal Online Gambling (IOG) Services in the Mobile Environment," *Security and Communication Networks*, vol. 2022, 2022, doi: 10.1155/2022/3286623.

[5] C. Wang, M. Zhang, F. Shi, P. Xue, and Y. Li, "A Hybrid Multimodal Data Fusion-Based Method for Identifying Gambling Websites," *Electronics (Switzerland)*, vol. 11, no. 16, Aug. 2022, doi: 10.3390/electronics11162489.

[6] Y. Zhang, X. Fu, R. Yang, and Y. Li, "DRSDetector: Detecting Gambling Websites by Multi-level Feature Fusion," in *Proceedings - IEEE Symposium on Computers and Communications*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1441–1447. doi: 10.1109/ISCC58397.2023.10217923.

[7] M. Aledhari, “Protecting Internet Traffic: Security Challenges and Solutions,” 2017.

[8] B. R. Dawadi, B. Adhikari, and D. K. Srivastava, “Deep Learning Technique-Enabled Web Application Firewall for the Detection of Web Attacks †,” *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042073.

[9] M. Tarigan, F. L. Gaol, and T. Matsuo, “Web Scraping Methods on Odoo Framework to Collect Rupiah Exchange Rate From Bank Indonesia Website,” in *Proceedings - 2021 10th International Congress on Advanced Applied Informatics, IIAI-AAI 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 697–702. doi: 10.1109/IIAI-AAI53430.2021.00124.

[10] C. Bhatt, Gaitri, D. Kumar, R. Chauhan, A. Vishvakarma, and T. Singh, “Web Scraping: Huge Data Collection from Web,” in *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology, ICSEIET 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 375–378. doi: 10.1109/ICSEIET58677.2023.10303037.

[11] A. Abodayeh, R. Hejazi, W. Najjar, L. Shihadeh, and R. Latif, “Web Scraping for Data Analytics: A BeautifulSoup Implementation,” in *Proceedings - 2023 6th International Conference of Women in Data Science at Prince Sultan University, WiDS-PSU 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 65–69. doi: 10.1109/WiDS-PSU57071.2023.00025.

[12] E. Smith, S. Michalski, K. H. K. Knauth, K. Kaspar, N. Reiter, and J. Peters, “Large-Scale Web Scraping for Problem Gambling Research: A Case Study of COVID-19 Lockdown Effects in Germany,” *J Gambl Stud*, vol. 39, no. 3, pp. 1487–1504, Sep. 2023, doi: 10.1007/s10899-023-10187-1.

[13] B. Jonas, F. Leuschner, A. Eiling, C. Schoelen, R. Soellner, and P. Tossmann, “Web-Based Intervention and Email-Counseling for Problem Gamblers: Results of a Randomized Controlled Trial,” *J Gambl Stud*, vol. 36, no. 4, pp. 1341–1358, Dec. 2020, doi: 10.1007/s10899-019-09883-8.

[14] B. R. Kamalov and M. V. Tumbinskaya, “Software for detecting ‘hidden miners’ in a browser environment,” *Journal Of Applied Informatics*, vol. 18, no. 1, pp. 96–110, Feb. 2023, doi: 10.37791/2687-0649-2023-18-1-96-110.

[15] L. Wang, Z. Zhang, and F. Ye, “Pornographic and Gambling Domain Recognition Method based on Long Distance Spare Multi-Head Self-Attention Vision-and-Language Model,” in *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications, AEECA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 504–508. doi: 10.1109/AEECA55500.2022.9919063.

[16] 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 (Switzerland)*, vol. 20, no. 14, pp. 1–21, Jul. 2020, doi: 10.3390/s20143989.

[17] D. Arunyagool, K. Chamnongthai, and D. Khawparisuth, “Monitoring and Energy Control Inside Home Using Google Sheets with Line Notification,” in *2021 International Conference on Power, Energy and Innovations, ICPEI 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 99–102. doi: 10.1109/ICPEI52436.2021.9690648.

[18] N. Chumuang, S. Hiranchan, M. Ketcham, W. Yimyam, P. Pramkeaw, and S. Tangwannawit, “Developed Credit Card Fraud Detection Alert Systems via Notification of LINE Application,” in *Proceedings - 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing, iSAI-NLP 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/iSAI-NLP51646.2020.9376829.

[19] W. Boonsong, “Smart Intruder Notifying System Using NETPIE through Line Bot Based on Internet of Things Platform,” in *2019 IEEE 5th International Conference on Computer and Communications (ICCC).*, IEEE, 2019.

[20] K. Sellamuthu, S. Ranjithkumar, K. Kavitha, and S. Gowtham, “On Page SEO Techniques for Better Ranking in Search Engines,” in *8th International Conference on Smart Structures and Systems, ICSSS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICSSS54381.2022.9782182.

[21] A. Nadeem, M. Hussain, and A. Iftikhar, “New Technique to Rank without off Page Search Engine Optimization,” in *Proceedings - 2020 23rd IEEE International Multi-Topic Conference, INMIC 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/INMIC50486.2020.9318166.

[22] R. Mitchell, *Web Scraping with Python, 2nd Edition*. 2018.

[23] P. Gong and S. Li, “Adaptive Behavior-aware Driven Intelligent Method for Detecting News Webpage Structure,” in *Proceedings - 2023 8th International Conference on Information Systems Engineering, ICISE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 115–120. doi: 10.1109/ICISE60366.2023.00030.

[24] M. Lutz, *Python Pocket Reference: Python in Your Pocket*. 2014.

[25] P. Tipauksorn, P. Luekhong, J. Thongpron, U. Kamnarn, K. Yingkayun, and A. Namin, “Stereo Vision-based Turn-Alignment Optimization for Wireless Power Transmission Positioning,” *2023 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*, pp. 1–8, 2023, doi: 10.1109/ITECAsia-Pacific59272.2023.10372364.

[26] M. Lutz, *Learning Python*, vol. 78, no. 1. 2007. doi: 10.1016/0019-1035(89)90077-8.

[27] “Higher education information dissemination system.” Accessed: Apr. 06, 2024. [Online]. Available: https://info.mhesi.go.th/homestat_academy.php

[28] “Casino Keywords - Find SEO & Google AdWords Key Words for Your Website.” Accessed: Apr. 06, 2024. [Online]. Available: https://www.wordstream.com/popular-keywords/casino-keywords

[29] “Top 50 SEO Keywords for Casinos | MarketKeep.” Accessed: Apr. 06, 2024. [Online]. Available: https://marketkeep.com/seo-keywords-for-casinos/

[30] O. Sakaliuk and O. Shershun, “INCREASE WEBSITE VISIBILITY ON THE INTERNET BY GOOGLE SEARCH CONSOLE,” *Automation of technological and business processes*, vol. 15, no. 2, pp. 12–17, Jun. 2023, doi: 10.15673/atbp.v15i2.2516.