Gambling Web Defacement Detection system (GWDD) in Chiang Mai University Domain

Department of Computer Science, Faculty of Science, Chiang Mai University, Thailand
204496 Cooperative Education, Year 2025

Name: Mr. Naruebordhin Pakwan 650510667 **Advisor:** Dr. Worawut Srisukkham
Department: Cyber Security Department (CSD), Information Technology Service Center (ITSC), Chiang Mai University

1 ABSTRACT

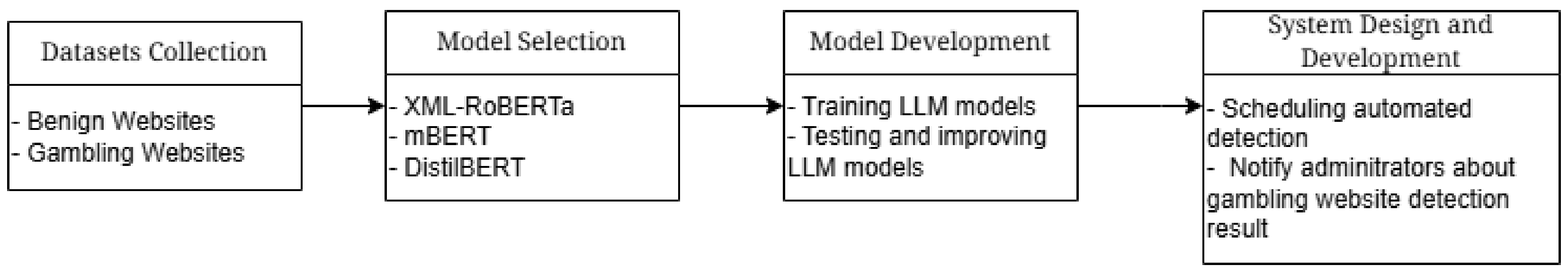
In recent years, website attacks in the form of embedded links or scripts from online gambling platforms have been steadily increasing, particularly within educational institution websites and the detection process of such embedded gambling content is often time-consuming. Therefore, this project aims to develop a detection system for gambling website embedding within the domain of Chiang Mai University to identify and resolve gambling-related issues more efficiently. The proposed approach employs a Large Language Model, XLM-RoBERTa, trained on two datasets: gambling-related content embedded within websites under educational domains in Thailand, and normal content collected from websites under the Chiang Mai University domain. For the prediction stage, the trained XLM-RoBERTa model was combined with a Mean-Pooling and Rule-Based method. The experimental results indicate that XLM-RoBERTa with a Mean-Pooling combined with Rule-Based prediction method achieves high performance, with an accuracy of 0.9937 on balanced data, a F1-score of 0.9592 on unbalanced data, and a F1-score of 0.9333 on context-based data. Finally, it can be concluded that XLM-RoBERTa combined with Mean-Pooling and Rule-Based prediction is highly effective for detecting gambling websites embedding within Chiang Mai University domain.

2 INTRODUCTION

Currently, many educational institutions in Thailand are facing a growing number of cybersecurity challenges. One of the most significant issues affecting these institutions is the defacement of departmental websites through the injection of code or scripts related to online gambling websites (web defacement). Chiang Mai University, which operates more than 2,600 domains under the Chiang Mai University domain has become a prominent target of such gambling-related web injections. This is because successfully embedding gambling content on Chiang Mai University websites increases the visibility of these gambling platforms and allows them to more easily reach their target audiences through various search engines.

Therefore, the ability to rapidly detect embedded gambling-related content can help reduce reputational damage, maintain user trust in Chiang Mai University, and enable faster remediation of compromised websites. Although tools for detecting gambling websites currently exist, most of them rely on search engine results rather than inspecting the actual content within websites. As a result, such approaches may lead to inaccurate or incomplete detection.

3 METHODOLOGY



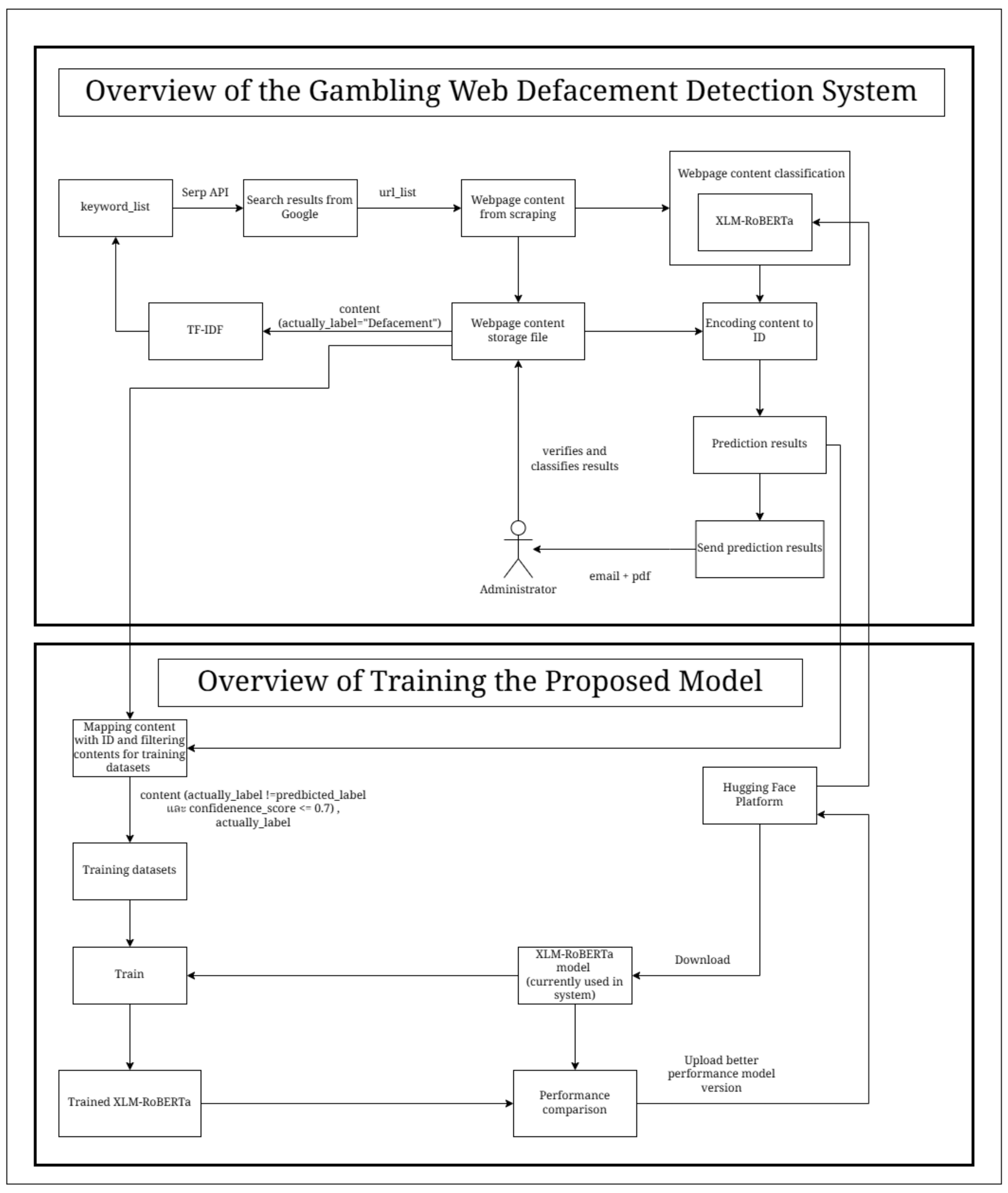
Datasets Collection
- Benign Websites
- Gambling Websites

Model Selection
- XML-RoBERTa
- mBERT
- DistilBERT

Model Development
- Training LLM models
- Testing and improving LLM models

System Design and Development
- Scheduling automated detection
- Notify adminitrators about gambling website detection result

4 OVERVIEW



Overview of the Gambling Web Defacement Detection System

Overview of Training the Proposed Model

keyword_list
Serp API
Search results from Google
url_list
Webpage content from scraping
Webpage content classification
XLM-RoBERTa
TF-IDF
content (actually_label="Defacement")
Webpage content storage file
Encoding content to ID
Prediction results
verifies and classifies results
Send prediction results
email + pdf
Administrator
Mapping content with ID and filtering contents for training datasets
content (actually_label !=predbicted_label และ confidenence_score <= 0.7) , actually_label
Hugging Face Platform
Training datasets
Train
XLM-RoBERTa model (currently used in system)
Download
Upload better performance model version
Trained XLM-RoBERTa
Performance comparison

6 CONCLUSION

Based on the performance evaluation results of the proposed models and prediction methods, it can be concluded that the XLM-RoBERTa model demonstrates strong performance and is well suited for detecting gambling-related content injections within websites under the domain of Chiang Mai University. For Gambling Web Defacement Detection system (GWDD), that was developed in the form of a command-based tool capable of operating on a scheduled basis within a Linux-based virtual machine environment. This system is able to automatically deliver detection results to relevant administrators. Although this system is capable of notifying administrators of detection results, certain limitations and errors remain and require further improvement.

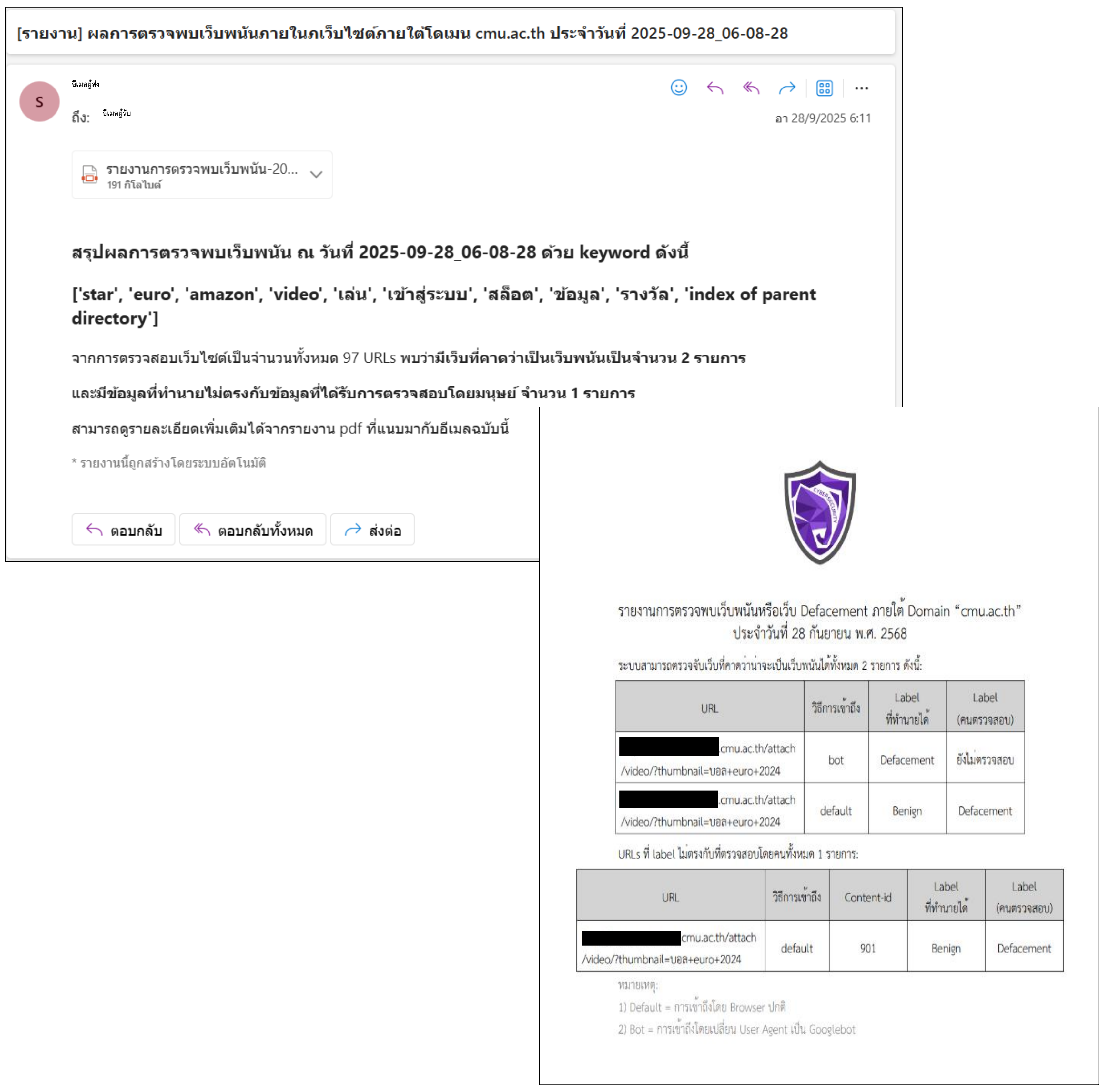
5 RESULTS

Table 1. F-1 score comparison of different prediction methods across balanced, unbalanced, and contextualized datasets.

Dataset Types/ Prediction methods	Balanced dataset	Unbalanced dataset	Contextualized dataset
Majority voting	0.9555	0.9388	1
Max-Pooling	0.9451	0.9406	1
Mean-Pooling with Rule-Based	0.9937	0.9542	0.9333

Table 2. F-1 score comparison of different models across balanced, unbalanced, and contextualized datasets.

Dataset Type/ Model	Balanced dataset	Unbalanced dataset	Contextualized dataset
TF-IDF with SVM	0.9885	0.9231	1
mBERT	0.9281	1	1
DistilBERT	0.8884	1	1
XLM-RoBERTa	0.9592	0.9333	0.9333



[รายงาน] ผลการตรวจพบเว็บพนันภายในภเว็บไซต์ภายใต้โดเมน cmu.ac.th ประจำวันที่ 2025-09-28_06-08-28

อีเมลผู้ส่ง
ถึง: อีเมลผู้รับ
อา 28/9/2025 6:11

รายงานการตรวจพบเว็บพนัน-20...
191 กิโลไบต์

สรุปผลการตรวจพบเว็บพนัน ณ วันที่ 2025-09-28_06-08-28 ด้วย keyword ดังนี้

['star', 'euro', 'amazon', 'video', 'เล่น', 'เข้าสู่ระบบ', 'สล็อต', 'ข้อมูล', 'รางวัล', 'index of parent directory']

จากการตรวจสอบเว็บไซต์เป็นจำนวนทั้งหมด 97 URLs พบว่า**มีเว็บที่คาดว่าเป็นเว็บพนันเป็นจำนวน 2 รายการ**

และมีข้อมูลที่ทำนายไม่ตรงกับข้อมูลที่ได้รับการตรวจสอบโดยมนุษย์ จำนวน 1 รายการ

สามารถดูรายละเอียดเพิ่มเติมได้จากรายงาน pdf ที่แนบมากับอีเมลฉบับนี้

* รายงานนี้ถูกสร้างโดยระบบอัตโนมัติ

รายงานการตรวจพบเว็บพนันหรือเว็บ Defacement ภายใต้ Domain “cmu.ac.th”
ประจำวันที่ 28 กันยายน พ.ศ. 2568

ระบบสามารถตรวจจับเว็บที่คาดว่าน่าจะเป็นเว็บพนันได้ทั้งหมด 2 รายการ ดังนี้:

URL	วิธีการเข้าถึง	Label ที่ทำนายได้	Label (คนตรวจสอบ)
[redacted].cmu.ac.th/attach /video/?thumbnail=บอล+euro+2024	bot	Defacement	ยังไม่ตรวจสอบ
[redacted].cmu.ac.th/attach /video/?thumbnail=บอล+euro+2024	default	Benign	Defacement

URLs ที่ label ไม่ตรงกับที่ตรวจสอบโดยคนทั้งหมด 1 รายการ:

URL	วิธีการเข้าถึง	Content-id	Label ที่ทำนายได้	Label (คนตรวจสอบ)
[redacted]cmu.ac.th/attach /video/?thumbnail=บอล+euro+2024	default	901	Benign	Defacement

หมายเหตุ:
1) Default = การเข้าถึงโดย Browser ปกติ
2) Bot = การเข้าถึงโดยเปลี่ยน User Agent เป็น Googlebot