2024 DUBLIN

2 - 4 October, 2024 / Dublin, Ireland

MULTIMODAL AI: THE SIXTH SENSE FOR CYBER DEFENCE

Younghoo Lee

Sophos, Australia

younghoo.lee@sophos.com

ABSTRACT

Given the growing complexity and diversity of cyber threats across various media formats, there is a pressing need to enhance cybersecurity defence mechanisms. This paper explores the potential of multimodal AI in addressing these evolving threats. Specifically, we analyse the use of multimodal AI for detecting phishing emails and classifying NSFW (Not Safe For Work) websites. By integrating and analysing both textual and visual data, multimodal AI can uncover hidden intentions that traditional methods might miss. Additionally, the analysed data supports classification predictions, which traditional machine learning lacks. We examine LLM-based multimodal AI and embedding models, demonstrating their effectiveness in processing diverse data modalities to deliver superior classification performance. Our experimental results show that multimodal AI significantly outperforms traditional machine learning approaches in both phishing detection and NSFW classification while providing added explainability, thus offering a more robust and effective defence mechanism for today's cybersecurity landscape.

1. INTRODUCTION

As cyber threats become increasingly complex and leverage diverse media formats, there is an urgent need for innovative cybersecurity defence mechanisms. This paper investigates the capabilities of multimodal AI, a sophisticated tool that transcends traditional text-based analysis by integrating data from text and images to uncover hidden threats. We evaluate the effectiveness of multimodal AI in detecting phishing emails and categorizing NSFW websites.

1.1 Multimodal AI for email phishing detection

Phishing emails often contain urgent prompts to verify delivery details or address payment discrepancies and typically include malicious URLs aimed at capturing user credentials. They closely mimic legitimate communications, making detection difficult for conventional spam filters. Although attackers may use generative multimodal AI to fabricate convincing phishing emails, the same AI technology can be harnessed to detect these fraudulent communications. Traditional machine learning (ML) models undergo phases like data collection, feature extraction (often using TF-IDF [Term Frequency-Inverse Document Frequency]), and model training and deployment. However, models like Random Forest or XGBoost, which rely on familiar words from training datasets, may overlook new phishing formats. Multimodal AI like GPT-4o from *OpenAI* or Gemini from *Google*, trained on diverse datasets, can identify phishing attempts by analysing email headers, content, and visual elements such as logos and images. These generative AI models can recognize new phishing emails efficiently, even without specific training data.

1.2 Embeddings as machine learning features for website classification

NSFW classification involves categorizing websites into groups like gambling, weapons, sports and games, with categories like gambling and weapons being NSFW. In ML, embeddings transform words or phrases into numerical vectors that capture meanings and relationships. Traditional approaches involve data collection, HTML content extraction, and feature creation using techniques like TF-IDF. However, these methods are limited by their inability to interpret non-text objects within images. Multimodal AI addresses this limitation by generating detailed descriptions of screenshots, capturing both textual and visual information, thus improving classification accuracy by providing a holistic representation of the website.

By leveraging the strengths of multimodal AI and advanced embedding techniques, our approach offers a robust solution to modern cybersecurity challenges. In the following sections, we will explore the details of our methodology, provide experimental validation, and discuss case studies that illustrate the efficacy of multimodal AI in both phishing detection and NSFW website classification.

2. DETECTING PHISHING EMAILS

This section highlights the limitations of current text-based phishing detection methods and demonstrates how multimodal AI decodes suspicious signals in email text and visual elements.

2.1 Challenges in detecting phishing emails

Phishing emails often include urgent requests to verify details or resolve payment issues and embed URLs for credential harvesting. Attackers craft these emails to resemble legitimate communications, incorporating brand images and genuine text while replacing legitimate URLs with malicious ones.

Traditional machine learning methods involve several steps. Initially, during data collection and feature extraction, ML models require extensive datasets for effective learning. Features such as text tokens (words) are extracted, and techniques like TF-IDF are used to assign importance to words. For instance, terms like ‘urgent’, ‘payment’, and ‘verify’ may carry higher TF-IDF scores due to their frequent occurrence in phishing emails. Then, during the training and evaluation phases, the dataset is split into training and test sets, with the model learning from the training data and being validated against the test data. Finally, the deployment phase entails monitoring incoming emails using the trained model. However, traditional models like Random Forest or XGBoost are limited because they depend on familiar words from the training dataset and may overlook new phishing emails that utilize unfamiliar words or images. For example, if a model lacks samples targeting the customers of Australian banks, it will likely miss phishing emails aimed at them. Models like Random Forest or XGBoost trained with TF-IDF features rely on frequent words in the training dataset to identify suspicious emails by detecting known combinations of suspicious words.

Multimodal AI based on large language models (LLMs), trained on massive datasets spanning numerous websites and brands, can identify phishing attempts by scrutinizing email headers, content, and visual elements such as logos and images. With specific examples and detailed instructions, generative models can adapt to recognize new and previously unseen phishing emails efficiently. Even without specific training data on novel phishing formats, these models can detect suspicious signals and the intent behind the email.

2.2 Leveraging multimodal AI for detection

Existing security procedures involve manual examination of suspicious emails for any signs of impersonation or malicious intent, looking at elements such as email headers, body content, and various UI features like brand logos. This process also includes analysing sender domains and URLs. These labour-intensive procedures can now be automated using large language models.

LLMs, trained on vast and diverse datasets, are capable of analysing suspicious elements within an email, understanding its intent, recognizing brand identities, and identifying impersonation attempts. They can do this by examining both the content and visual aspects of an email. These models can identify dubious domains or URLs without needing additional training or external databases. With just a few examples, they can detect new phishing emails, a feat that traditional machine learning methods can only achieve with large-scale training datasets.

We use GPT-4o, a multimodal AI from *OpenAI*, to inspect emails as part of our phishing detection methodology. We provide the LLM with specific instructions to uncover suspicious elements from the email data and screenshot images of the email or its URLs, which are generated by rendering the email's HTML content. The report that we get from our instructions provides a classification score indicating the presence of any suspicious elements. Such a report can reveal the suspicious nature of an email sender's display name or domain and any impersonation attempts in the URLs or images, something that traditional methods are unable to provide.

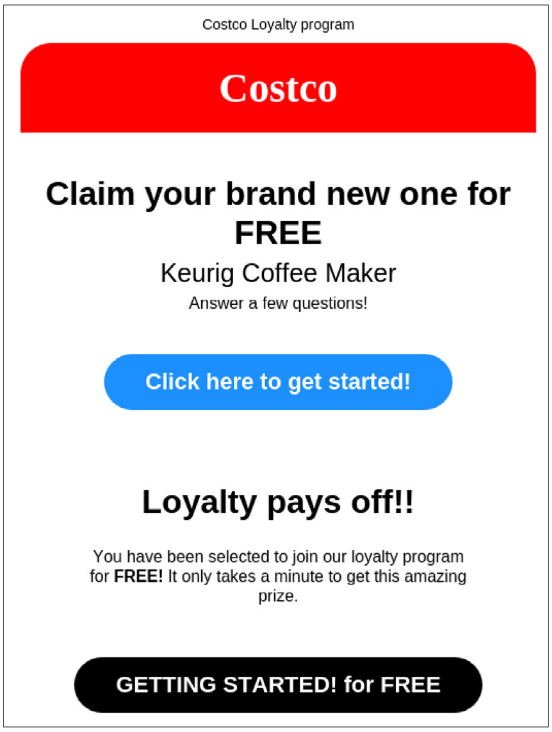
The GPT-4o instruction for phishing detection:

> Analyse the email data and the accompanying screenshot images if available and generate a report in the following JSON format to detect and quantify the level of suspiciousness in the email, considering common indicators of phishing, fraud, or malicious intent.
>
> - 'summary': A brief overview of the content of the email.
> - 'suspicious_elements_domain_content': Review the sender's email domain for consistency with the content of the email. Look for domain spoofing or the use of domains similar to reputable domains to mislead the recipient.
> - 'suspicious_text': Highlight text in the email that indicates a sense of urgency, incites immediate action, or otherwise seems intended to manipulate the recipient emotionally.
> - 'suspicious_links': Catalog top 3 links found in the email and assess each for potential malicious intent, especially those directing to suspicious or misspelled domains.
> - 'suspicious_images': Analyse the accompanying screenshots for any indicators of phishing.
> - 'impersonated_target_in_image': The impersonated brand or target in the images, the sender's domain does not match with the target.
> - 'suspicious_score': Provide an overall score between 0.0 (not suspicious) and 1.0 (extremely suspicious) based on the aggregated suspicious indicators found in the email.

2.3 Case studies

Case: Costco

Our approach identified a phishing email targeting *Costco* clients. The method detected suspicious signals in the email header and content, resulting in a suspicious score of 0.8 (on a scale from 0.0 to 1.0). When the screenshot image was analysed with GPT-4o, the multimodal AI recognized additional suspicious signals from both the email data and image, increasing the suspicious score to 0.9 – sufficient to mark it as malicious.



Costco Loyalty program

Costco

Claim your brand new one for FREE

Keurig Coffee Maker

Answer a few questions!

Loyalty pays off!!

You have been selected to join our loyalty program for **FREE!** It only takes a minute to get this amazing prize.

Figure 1: Costco phishing email.

The following GTP-4o report for the Costco email identifies the sender’s suspicious domain and intention:

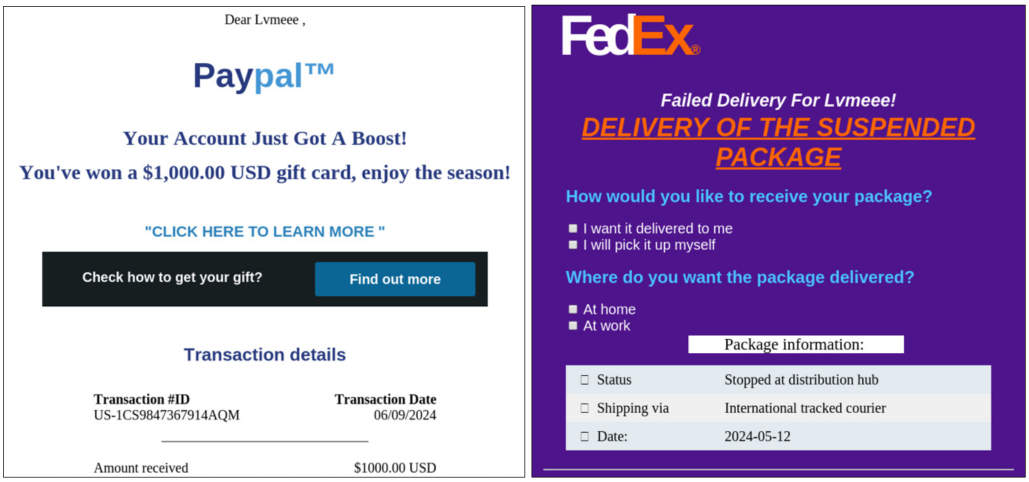
```
{
  "subject": "RE:You have won a Keurig Coffee Maker. #ID84656",
  "summary": "Email claiming the recipient has won a Keurig Coffee Maker and prompts them to click a link to claim the prize.",
  "sender": "Costco® <teamsupport-Lvmeee-923309212439507@ocxpnbnjfwuu.com.au>",
  "suspicious_elements_domain_content": "The domain 'ocxpnbnjfwuu.com.au' is not consistent with legitimate Costco domains, indicating potential domain spoofing.",
  "suspicious_elements_links_content": "The link 'https://storage.googleapis.com/' does not align with Costco's official URLs and uses a Google Cloud Storage domain to obscure the actual destination.",
  "suspicious_text": "Phrases like 'Claim your brand new one for FREE!' and 'You have been selected to join our loyalty program for FREE!' create a sense of urgency and pressure the recipient to take immediate action.",
  "impersonated_target_in_text": "Costco",
  "suspicious_score": 0.8
}
```

The following GPT-4o report for the Costco email and its screenshot image recognizes the additional impersonation attempt in visual elements.

```
{
…
  "suspicious_images": "The screenshot images contain visual elements that mimic genuine Costco branding. There are call-to-action buttons to claim a prize, which is a common phishing tactic. The URLs do not match with the sender’s legitimate domain.",
  "impersonated_target_in_image": "Costco",
  "suspicious_score": 0.9
}
```

Case: Other brands

Similarly, our approach detected phishing emails impersonating *PayPal* and *FedEx*. The method identified the suspicious domain and UI elements within both the email and its screenshot image.



Dear Lvmeee ,

Paypal™

Your Account Just Got A Boost!

You've won a $1,000.00 USD gift card, enjoy the season!

"CLICK HERE TO LEARN MORE "

Transaction details

Transaction #ID	Transaction Date
US-1CS9847367914AQM	06/09/2024
Amount received	$1000.00 USD

Failed Delivery For Lvmeee!

DELIVERY OF THE SUSPENDED PACKAGE

How would you like to receive your package?

- I want it delivered to me
- I will pick it up myself

Where do you want the package delivered?

- At home
- At work

Package information:

☐ Status	Stopped at distribution hub
☐ Shipping via	International tracked courier
☐ Date:	2024-05-12

Figure 2: PayPal and FedEx phishing emails.

2.4 Experiment results

We conducted an experiment to compare the effectiveness of our phishing detection approach against baseline machine learning models: Random Forest and XGBoost. These models utilized TF-IDF features derived from raw email data. The dataset, consisting of approximately 2,000 balanced samples, was split into 70% for training and 30% for testing in a random split experiment. Additionally, to evaluate the models' performance on unseen samples, we conducted a brand split experiment by training on emails impersonating certain brands (e.g. *Google* and *Microsoft*) and testing on emails targeting other brands (e.g. *PayPal* and *Costco*).

The results, which are summarized in Tables 1 and 2, indicate the following:

- When evaluated on randomly split data, both Random Forest and XGBoost demonstrated high performance with an F1 score close to 0.991.
- On unseen samples, however, their performance dropped significantly, with Random Forest achieving an F1 score of 0.529 and XGBoost slightly better at 0.657. This indicates that these models struggled to generalize effectively to new threats during the brand split experiment.

Random Forest	**Random split**	**Brand split**
Precision	1.000	0.583
Recall	0.983	0.401
F1	0.991	0.529

Table 1: Performance of Random Forest with TF-IDF features.

XGBoost	**Random split**	**Brand split**
Precision	1.000	0.583
Recall	0.966	0.752
F1	0.983	0.657

Table 2: Performance of XGBoost with TF-IDF features.

Our approach with GPT-4o, which integrates both textual and image data descriptions for the test dataset in the brand split, significantly outperformed the baseline models. Through the fusion of features from both text and images, we were able to achieve superior detection performance against unseen threats. The results are presented in Table 3.

GPT-4o	**Email text input**	**Email text and image input**
Precision	1.000	0.982
Recall	0.936	0.961
F1	0.967	**0.971**

Table 3: The performance of GPT-4o for brand split with text and image input.

In conclusion, our multimodal AI approach, which leverages both text and image inputs, offers a more robust solution for detecting phishing attempts, particularly when facing unseen threats. The use of both text and image features proved to be more effective, as reflected by the highest F1 score of 0.971 in the brand split test set.

3. CLASSIFYING NSFW WEBSITES

The proliferation of NSFW content on the internet poses a significant challenge for web content classification. Embedding models powered by multimodal AI interpret the 'language' of images, facilitating the robust and accurate classification of NSFW websites. This section outlines the application of embeddings for NSFW classification, encompassing categories like gambling, weapons, sports and games, with gambling and weapons classified as NSFW.

In machine learning, embeddings are numerical representations of data. For text data, embeddings convert words or phrases into numerical vectors that capture their meanings, similarities and interrelations. For instance, the words 'gambling' and 'casino' would be proximate in the embedding space, indicating their similarity. In a traditional approach, initial steps involve data collection, extracting HTML content from URLs, and tokenizing words. Features like TF-IDF are used to represent the text numerically. For instance, terms such as 'bet', 'poker' and 'jackpot' would have higher TF-IDF scores on a gambling site. Additional features can be derived using OCR (Optical Character Recognition) to extract text from web page images. However, a limitation of traditional OCR is its inability to interpret non-text objects, missing vital details embedded within images.

3.1 Leveraging contextual embedding

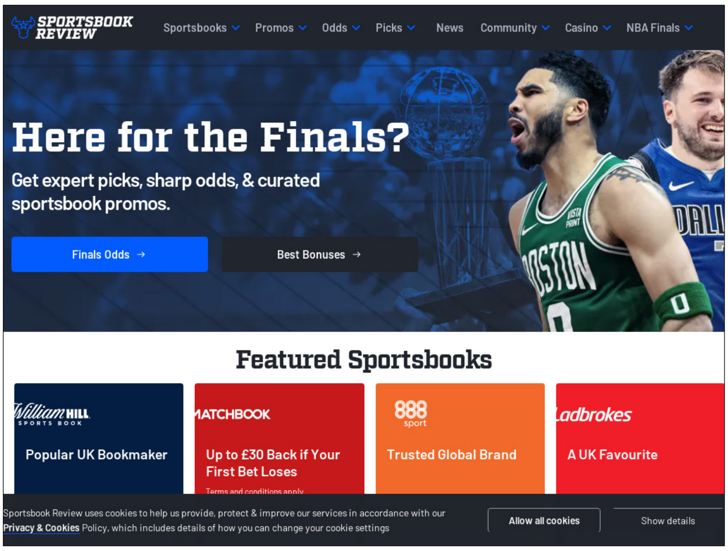
Leveraging multimodal AI, we can obtain detailed descriptions of screenshots, which are critical for identifying a website's purpose. This approach addresses the limitation of traditional text analysis by including visual context. For instance, it can describe text as well as visual elements like poker chips or roulette tables on a gambling site, providing additional explainability for our classification problem. The textual description of the HTML content and the screenshot can be used as input to generate TF-IDF features in the traditional ML pipeline. Furthermore, these text descriptions can be transformed into more representative numerical vectors using LLM-based embeddings. These contextual embedding vectors help to identify similar websites and can be used as features to train traditional models like Random Forest or XGBoost.

We employ GPT-4o to generate detailed descriptions of the HTML content and its corresponding screenshots. *OpenAI*'s embedding model, text-embedding-3-small, is then used to create embedding vectors from these descriptive texts. These embeddings are subsequently used to train ML models, such as Random Forest and XGBoost, for the web categorization task. Our experiments show that those models trained with embedding features significantly outperform those using traditional word-based TF-IDF features.

3.2 Case studies

Case: Gambling website

A gambling website was misclassified as a sports site by a Random Forest model trained with TF-IDF features from raw HTML data. This misclassification likely occurred due to the sports-related content present on the site. However, our approach, which includes descriptions from both HTML content and images, enabled the same Random Forest model to correctly classify it as a gambling site.



Here for the Finals?

Get expert picks, sharp odds, & curated sportsbook promos.

Featured Sportsbooks

Popular UK Bookmaker

Up to £30 Back if Your First Bet Loses

Terms and conditions apply

Trusted Global Brand

A UK Favourite

Sportsbook Review uses cookies to help us provide, protect & improve our services in accordance with our Privacy & Cookies Policy, which includes details of how you can change your cookie settings

Figure 3: Gambling website.

The following descriptions from GPT-4o demonstrate its ability to understand the content of multimodal AI. GPT-4o accurately described the web page and provides relevant details for the task.

The description from the screenshot:

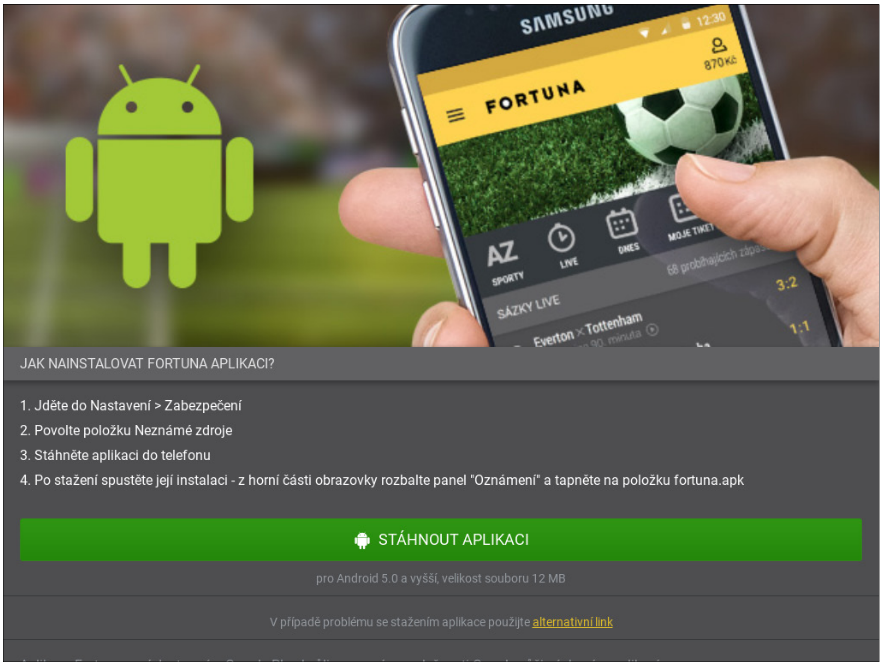
> The screenshot image contains the following main objects:
>
> 1. Two basketball players in sports jerseys.
> 2. A large trophy in the background.
> 3. Text phrases such as ‘Here for the Finals?’, ‘Get expert picks, sharp odds, & curated sportsbook promos’, ‘Finals Odds’, ‘Best Bonuses’, and ‘Featured Sportsbooks’.
> 4. Logos and names of various sportsbooks (e.g. William Hill, Matchbook, 888sport, Ladbrokes).

The description from the HTML data:

> Here are the main features:
>
> 1. Title:
>   - The title of the page is ‘Sportsbook Review | SBR - Sports Betting Experts since 1999’.
> 2. Keywords and Content:
>   - The keywords and phrases such as ‘sports betting’, ‘odds’, ‘picks’, ‘reviews’, ‘betting community’, ‘sportsbooks’, ‘sports betting apps’, ‘sports betting promos’, ‘sports betting landscape’, and ‘sports betting bonuses’ are repeated multiple times throughout the HTML content.

Case: non-English gambling website

A particular website was initially categorized as a sports site using the Random Forest model trained with raw HTML data. However, when the Random Forest model was trained with embeddings, it accurately identified the site as a gambling site. Notably, GPT-4o captured all relevant details, even though the site content was not in English. The generated description stated, ‘Text and icons related to sports, such as “SPORTY” and “SÁZKY LIVE”, and scores of soccer matches (e.g. Everton vs. Tottenham). These elements indicate that the website is related to sports, specifically soccer, and potentially sports betting due to the presence of match scores and the context of the app being promoted.’



JAK NAINSTALOVAT FORTUNA APLIKACI?

1. Jděte do Nastavení > Zabezpečení
2. Povolte položku Neznámé zdroje
3. Stáhněte aplikaci do telefonu
4. Po stažení spustěte její instalaci - z horní části obrazovky rozbalte panel "Oznámení" a tapněte na položku fortuna.apk

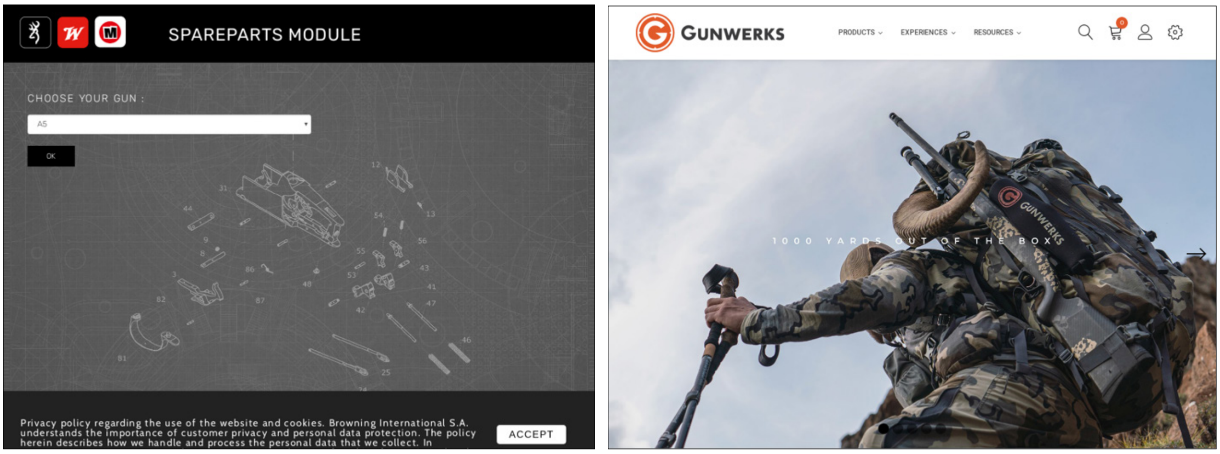
pro Android 5.0 a vyšší, velikost souboru 12 MB

V případě problému se stažením aplikace použijte alternativní link

Figure 4: Non-English gambling website.

Case: Weapons website

Some websites with limited textual information were correctly classified as weapons sites due to the image-based information captured by multimodal AI.



Privacy policy regarding the use of the website and cookies. Browning International S.A. understands the importance of customer privacy and personal data protection. The policy herein describes how we handle and process the personal data that we collect. In

Figure 5: Weapons websites.

3.3 Experimental results

We conducted an experiment to compare our proposed approach with baseline models, specifically Random Forest and XGBoost models, utilizing TF-IDF features derived from raw HTML data and text description from GPT-4o. While it's feasible to use a vector database with embedding vectors for searching similar websites, we opted to train classification ML models with both word-based TF-IDF vectors and LLM-based embedding vectors to assess their performance in classification tasks.

The Random Forest model exhibited various levels of performance with different input features, as shown in Table 4. For the random split experiment, the dataset was divided into a 70% training set and a 30% test set; this dataset consisted of approximately 2,000 samples. To prevent overlap of similar websites, the dataset was deduplicated by domains. Initially, using TF-IDF features from raw HTML data, the model achieved an F1 score of 0.84. When TF-IDF features from HTML text descriptions generated by GPT-4.0 were utilized, the F1 score improved to 0.92. With LLM embedding features from the same text descriptions, the F1 score further increased to 0.94. The highest performance was noted when embeddings from both HTML content and screenshots were combined, resulting in an F1 score of 0.96. Thus, the inclusion of more detailed and combined feature embeddings significantly enhanced the model's classification performance. Table 5 shows similar trends for the XGBoost model.

Random Forest	**TF-IDF with raw HTML**	**TF-IDF with HTML text**	**Embedding with HTML text**	**Embedding with HTML and Image text**
Precision	0.891	0.939	0.958	0.970
Recall	0.813	0.912	0.930	0.951
F1	0.841	0.924	0.942	**0.960**

Table 4: The classification performance of Random Forest.

XGBoost	**TF-IDF with raw HTML**	**TF-IDF with HTML text**	**Embedding with HTML text**	**Embedding with HTML and Image text**
Precision	0.901	0.907	0.960	0.959
Recall	0.864	0.907	0.955	0.955
F1	0.880	0.907	0.958	**0.957**

Table 5: The classification performance of XGBoost.

These experimental results clearly demonstrate that embeddings derived from combined descriptions of HTML and images provide more efficient representations of websites. This, in turn, enhances the classification performance of general-purpose machine learning models, as evidenced by the improved precision, recall, and F1 scores across both Random Forest and XGBoost models.

4. CONCLUSION

Multimodal AI offers a powerful approach to navigating the increasingly complex world of cybersecurity threats. By merging data from various modalities, such as text and images, multimodal AI systems provide a more comprehensive understanding of data intent. Our research in phishing detection and NSFW website categorization shows that multimodal AI significantly outperforms traditional text-based machine learning techniques. Specifically, large language models (LLMs) based on multimodal AI and advanced embedding techniques exhibit superior classification capabilities, even against previously unseen threats, while also providing decision explainability. This research underlines the necessity of incorporating multimodal AI into cybersecurity structures for more precise and efficient threat detection and categorization. The concrete results detailed in this paper underscore the substantial advantages of multimodal AI, setting the stage for its wider use in protecting digital landscapes.

REFERENCES

[1] Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Amodei, D. Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165. 2020. https://arxiv.org/abs/2005.14165.

[2] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Bengio, Y. Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27, 2672-2680. 2014. https://papers.nips.cc/paper/5423-generative-adversarial-nets.

[3] Ho, T. K. Random Decision Forests. Proceedings of the Third International Conference on Document Analysis and Recognition, 278-282. 1995. https://ieeexplore.ieee.org/document/598994.

[4] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. 2016. https://dl.acm.org/doi/10.1145/2939672.2939785.

[5] Kowsari, K.; Meimandi, K. J.; Heidarysafa, M.; Mendu, S.; Barnes, L. E.; Brown, D. E. Text Classification Algorithms: A Survey. Information, 10(4), 150. 2019. https://www.mdpi.com/2078-2489/10/4/150.

[6] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. 2013. https://arxiv.org/abs/1301.3781.

[7] OpenAI. ChatGPT: Optimizing Language Models for Dialogue. 2023. https://openai.com/research/chatgpt.

[8] Raff, E.; Barker, J.; Sylvester, J.; Brandon, R.; Catanzaro, B. Malware Detection by Eating a Whole EXE. arXiv preprint arXiv:1710.09435. 2018. https://arxiv.org/abs/1710.09435.

[9] Saber, M. I.; Chakraborty, M.; Ghose, M. K. Detection of Phishing Emails Using Machine Learning and Deep Learning. Journal of King Saud University-Computer and Information Sciences, 33(9), 1166-1176. 2021. https://doi.org/10.1016/j.jksuci.2019.06.001.

[10] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Polosukhin, I. Attention is All you Need. Advances in Neural Information Processing Systems, 30, 5998-6008. 2017. https://papers.nips.cc/paper/7181-attention-is-all-you-need.

[11] Zhang, H.; Cisse, M.; Dauphin, Y. N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. International Conference on Learning Representations. 2018. https://arxiv.org/abs/1710.09412.