Large Language Models for Detecting Gambling Advertisement Images to Enhance the Efficiency of the Creative Review Process

Edward L Martis
Samsung Ads
Bengaluru, Karnataka, India
e.martis@samsung.com

Jayesh Santosh Asawa
Samsung Ads
Bengaluru, Karnataka, India
jayesh.asawa@samsung.com

ABSTRACT

The creative review process often involves manually sifting through numerous images to identify content that violates advertising policies. The manual creative review team spends a lot of time and resources in detecting harmful content, such as gambling within image creatives. Our objective is to reduce the workload of the team with advanced ML models by classifying the images as gambling or not. For this purpose, we have evaluated CNN based model (VGG-16), Vision Transformer model (ViT), and LLMs (LLAMA-Vision-11b and LLM2Vec encoder), along with an ensemble of these for the gambling image classification task. We are able to achieve close to 99% accuracy, with misclassification of gambling (FNR) and non-gambling (FPR) being 5% and 1%, respectively. We have also achieved an approximately 95% potential reduction in time, effort, and cost on application of our classification solution.

1 INTRODUCTION

Recent discussions on responsible gambling policies highlight a concerning rise in gambling disorders among adolescents, underscoring the negative effects of unrestricted access to addictive activities [1]. Protecting children and teenagers from harmful information is essential for their safety. This paper seeks to develop a method for filtering online images that contain potentially harmful content, such as gambling, and can potentially be extended to other sensitive categories. By implementing this method, it is possible to create services that safeguard adolescents while reducing their exposure to potentially dangerous content. The manual review of creative content (e.g., ads images) frequently necessitates the examination of a substantial volume of images to identify instances of advertising policy violations. In the context of gambling advertisements, this process can be particularly challenging due to the subtle nature of some promotional materials. Additionally, manual review plays a critical role in maintaining trust between brands and their audiences, as it ensures that only compliant and safe content is displayed to users, thereby preserving the brand reputation and user trust. Our review process uses widely available third-party APIs, such as cloud vision, for initial filtering. For specialized policies like rejecting gambling content such API's doesn't perform well. Nearly, 70% of gambling images were marked safe. As a result, the creative review team has to manually classifies each and every image, which is time-consuming, costly, and prone to human error. To overcome this challenge, we investigated and applied a combination of traditional CNNs, transformer models, and LLM2Vec models to achieve better classification performance compared to the third-party API. We targeted only gambling images for the first phase, where we could reduce the manual review effort in terms of turnaround time and cost. With the use of our new system, the manual review team will be able to spend time reviewing images more effectively, allowing them to focus on the most critical cases and improving overall workflow efficiency. We do this by providing a group of images where manual review is required, as shown in the results section (Table 2).

CCS CONCEPTS

• Computing methodologies → Artificial intelligence → Computer vision; • Computing methodologies → Machine Learning; • Information systems → Information retrieval → Retrieval models and ranking → Language models;

KEYWORDS

SSP, Creatives, Manual Review, Gambling, LLM, Prompt Engineering

2 RELATED WORKS

In this section, we briefly review some related works for gambling detection methods and approaches. In the domain of harmful content detection, particularly for gambling-related websites and images, several studies have contributed innovative approaches. Paper [2] investigates a visual-based approach to detect harmful

KDD '25, August, 2025, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/06/08
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference format:

Edward Martis and Jayesh Santosh Asawa 2025. Large Language Models for Detecting Gambling Advertisement Images to Enhance the Efficiency of the Creative Review Process. In Proceedings of (AdKDD'25). ACM, Toronto, Canada, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

gambling and pornographic websites, addressing the limitations of traditional text and URL-based methods due to rapid domain name changes. By utilizing the Bag-of-Words (BoW) model on website screenshots, the study successfully enhances classification efficiency, as demonstrated by promising experimental results on collected datasets.

Paper [3] presents an advanced approach to detect pornographic and gambling websites by integrating both visual and textual analysis. This method, known as PG-VTDM (Porn and Gambling Visual and Textual Decision Mechanism), utilizes Doc2Vec for textual feature extraction and improves the bag-of-visual-words (BoVW) technique by incorporating local spatial relationships for better visual feature representation. The system's decision mechanism employs logistic regression to fuse classification results from both text and image classifiers, achieving high accuracy metrics exceeding 99%. However, this paper focuses mainly on website detection which generally contains more than just an image, there can be text or multiple images hence it contains more information than just an image in contrast to our work. Specially the keywords within the website content helps in accuracy and other metrics a lot.

In Paper [4], the authors address the gap in harmful image recognition for gambling and drugs by introducing a comprehensive dataset encompassing images from domains like pornography, gambling, violence, and drugs. They have leveraged the CLIP (Contrastive Language-Image Pre-training) model with strongly and weakly associated prompts, achieving zero-shot learning with high accuracy and low learning costs. Specifically for the gambling use case, this approach demonstrates how vision-language models can effectively recognize harmful content, even without extensive labeled data, providing a practical solution for safeguarding children and vulnerable populations from exposure to gambling-related risks in digital environments.

Our main work builds on these studies by exploring the use of Transformers and Large Language Models (LLMs) for gambling detection and broader harmful content recognition. This approach not only integrates the strengths of visual and textual analysis but also leverages zero-shot learning to address data scarcity, particularly in niche areas like gambling detection. By doing so, we aim to enhance the robustness and adaptability of content detection systems, contributing to the evolution of detection methods in this field. To the best of author's knowledge, this is the first time exploration of application of LLM via prompt engineering in detection of gambling and potentially other harmful contents. With the boom of AI and LLM trained with such a huge amount of data, they have got smart enough to detect the harmful content and classify the images with great accuracy.

3 METHODOLOGY

3.1 Dataset

We created a specialized dataset tailored for the gambling-specific use case, derived from proprietary data logs of Samsung's Supply-Side Platform (SSP), which serves as ad inventory to various publishers. This dataset encompasses a wide array of gambling scenarios, including casino-style gambling (e.g., roulette, blackjack, slot machines), real money games (skill-based games with monetary rewards), cryptocurrency gambling (platforms utilizing digital currencies for wagering), lottery, and luck-based gambling advertisements (promotions for sweepstakes, raffles, etc.). Hence the dataset was general enough containing varieties of gambling and non-gambling images. Initially, the dataset comprised 500,000 images, collected over a defined period from our ad serving logs. Our manual review team, consisting of trained annotators, meticulously labelled each image with relevant categories such as "gambling", "games" (non-gambling related), and "cards" (playing cards in both gambling and non-gambling contexts), among other classifications to capture a comprehensive understanding of the image content. This initial labelling phase was crucial for establishing a ground truth for our classification task.

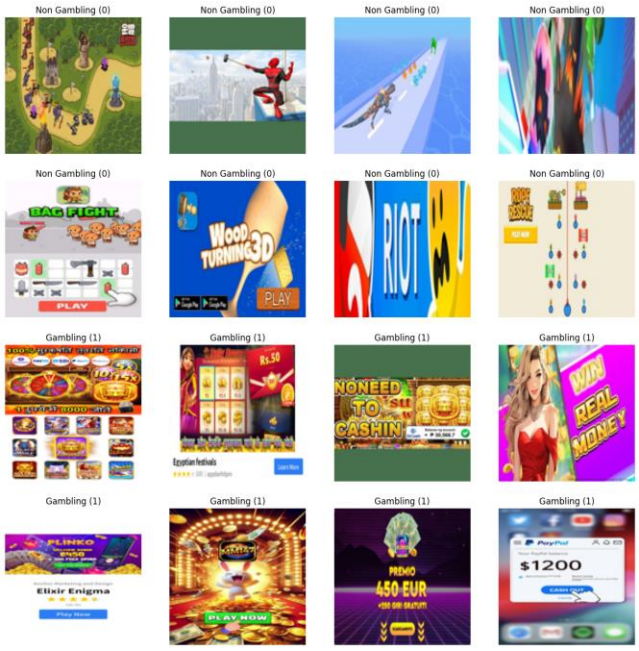

Fig. 1. Sample Creative Ads images in Dataset with Labels.

Upon initial inspection for data quality, a significant redundancy was identified, with 90% of the images being exact or near-duplicates. This high level of duplication could skew model training and evaluation, leading to overfitting and an inaccurate assessment of generalization performance. To address this, we employed a robust image hashing algorithm and a similarity threshold to identify and remove these duplicates. After this deduplication process, we were left with a more representative set of 50,000 unique images, ensuring a more balanced and informative dataset. The subsequent step involved the crucial label creation process for our binary classification task (gambling vs. non-gambling). We designated images tagged with "gambling" and "cards" (when contextually indicative of gambling) as label 1 (positive class, representing gambling content) and the remaining images, which included "games" and other non-gambling related

categories, as label 0 (negative class). This labelling strategy aimed to capture the core objective of identifying gambling-related visuals. Following this rigorous cleaning and labelling process, the refined dataset consisted of 30,000 images, providing a substantial volume for training our models. Notably, 4% of these images corresponded to the gambling class (label 1). Fig. 1. Shows the small sample of the 16 images from our dataset showcasing the images belonging to Gambling and Non-Gambling class. For evaluating the performance and generalizability of our classification models, we resized the images to standardized dimensions of 64x64 and 128x128 pixels. These resolutions were chosen as a trade-off between computational efficiency and the preservation of essential visual features. The final dataset was strategically split into training, testing, and validation sets in a 60:20:20 ratio, respectively. The training set (18,000 images) was used to learn the model parameters, the validation set (6,000 images) was used for hyperparameter tuning and early stopping during training to prevent overfitting, and the testing set (6,000 images), held out from the training process, was used for the final unbiased evaluation of the trained models' performance on unseen data.

3.2 Modeling Methods

We used a multi-pronged approach for classifying the images as gambling or not. Our strategy involves implementation of traditional CNN based model, followed by the vision transformer model, LLM based approaches and an ensemble approaches for selection of best approach in terms of performance.

3.2.1 Traditional CNN Model:

VGG-16 [5], a deep Convolutional Neural Network (CNN) architecture, is distinguished by its significant depth, comprising 13 convolutional layers and 3 fully connected layers, totalling 16 layers with learnable weights. This increased depth compared to earlier CNN architectures allows the model to learn a hierarchy of increasingly complex visual features from raw pixel inputs. The convolutional layers employ small 3x3 filters, enabling the network to capture fine-grained details in images. Multiple stacked convolutional layers with ReLU activation functions are followed by max-pooling layers to reduce spatial dimensions and introduce translational invariance. The subsequent fully connected layers learn high-level representations of the image content, culminating in a softmax output for classification. This deep architecture endows the model with a remarkable capacity to identify intricate features from images, making it particularly suitable for complex image classification tasks. The model's demonstrated success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) highlights its power in learning discriminative visual patterns, rendering it a strong baseline model for the challenging task of gambling image classification, where subtle visual cues can be critical.

The fact that VGG-16 is pre-trained on the massive ImageNet dataset, containing millions of diverse images spanning 1000 object categories, equips it with a robust foundation for recognizing a wide range of visual concepts. This pre-training process allows the model to learn generic image features that are transferable to other computer vision tasks. This pre-trained knowledge facilitates efficient fine-tuning for specific tasks, such as identifying gambling-related content, as the initial layers of the network have already learned useful visual representations. By leveraging transfer learning, we can adapt the pre-trained VGG-16 model to our smaller gambling dataset, achieving good performance with significantly less training data and computational resources compared to training a deep network from scratch.

In our implementation, we employed transfer learning techniques to adapt the VGG-16 model, pre-trained on ImageNet, to our specific gambling image classification task. We replaced the original classification layer of the VGG-16 model with a new fully connected layer with two output neurons, corresponding to the gambling (label 1) and non-gambling (label 0) classes, followed by a softmax activation function. The weights of the pre-trained convolutional layers were initially frozen to preserve the learned generic image features. The model was then trained using our gambling dataset with the Adam optimizer, a computationally efficient stochastic gradient descent algorithm, and a relatively small learning rate of 0.0001 to avoid destabilizing the pre-trained weights during fine-tuning. To address the significant class imbalance in our dataset, where the number of non-gambling images was much larger than gambling images, we applied class weights in a 1:40 ratio for the non-gambling (0) and gambling (1) labels, respectively. This weighting scheme assigns a higher penalty to misclassifying gambling images, forcing the model to pay more attention to the minority class. We initially trained the model for 100 epochs, monitoring the validation accuracy and loss to detect signs of overfitting. However, observations from Fig. 2, which plots the training and validation accuracy over epochs, revealed that the validation accuracy plateaued and even slightly decreased after approximately 40 epochs, indicating that further training was not leading to improved generalization and could be causing overfitting.

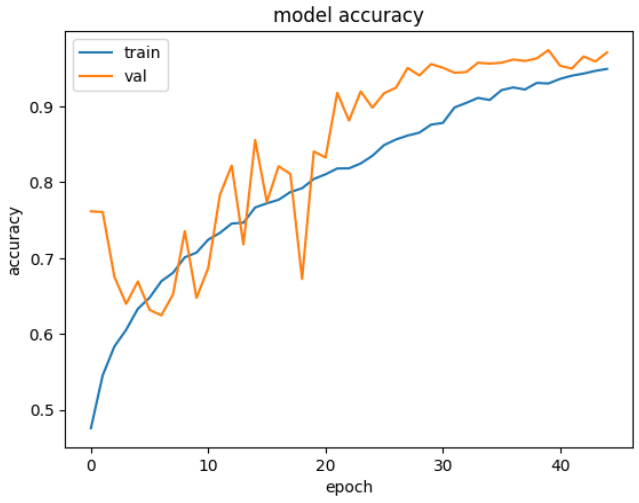

Fig. 2. Model Training for VGG-16 (with 40 epochs)

To further optimize the model and potentially improve performance, we iteratively experimented with freezing different numbers of the last few convolutional layers of the pre-trained VGG-16 model during subsequent training rounds. This allowed us to fine-tune different levels of the feature hierarchy learned by the pre-trained network. We also experimented with input image sizes of both 64x64 and 128x128 pixels to assess the impact of input resolution on model performance and computational cost. Model performance across different configurations and training strategies was rigorously compared using key evaluation metrics such as False Negative Rate (FNR), which measures the proportion of actual gambling images classified as non-gambling (a critical error in our use case), False Positive Rate (FPR), which measures the proportion of non-gambling images classified as gambling, and the F1-score, which provides a balanced measure of precision and recall on the test set, taking into account the class imbalance.

3.2.2 Transformer Model:

The Vision Transformer (ViT) [6] model represents a significant departure from traditional Convolutional Neural Networks (CNNs) in its approach to image processing, by leveraging the self-attention mechanisms that have proven highly successful in natural language processing. Unlike CNNs, which process images through localized receptive fields and a hierarchy of convolutional layers to capture spatial hierarchies, ViT treats an input image as a sequence of visual tokens, similar to words in a sentence. To achieve this, ViT first divides an input image into a fixed number of non-overlapping patches of a specified size (e.g., 16x16 pixels). Each patch is then flattened into a linear vector and passed through a linear embedding layer to project it into a high-dimensional space. To retain positional information, which is crucial for understanding the spatial arrangement of the patches, a positional encoding is added to each embedded patch. These embedded patches, now augmented with positional information, are then fed into a standard Transformer encoder architecture, consisting of multiple layers of multi-head self-attention and multi-layer perceptrons. The self-attention mechanism allows each image patch to attend to all other patches in the sequence, enabling the model to capture global relationships and long-range dependencies across the entire image, regardless of their spatial distance.This approach allows the model to focus on important parts of the image dynamically. ViT's flexibility in processing image patches is particularly advantageous for detecting gambling-related content, where important features may vary in appearance or positioning. This capability enables the model to adapt to different visual contexts effectively.

To assess the efficacy of the Vision Transformer (ViT) model for our specific classification task, we performed transfer learning by fine-tuning a pre-trained ViT architecture using our dataset. Our initial training phase spanned 100 epochs, aiming to fully adapt the model to our data distribution. Analysis of the training and validation curves, particularly illustrated in Fig 3, revealed a plateau in validation accuracy after approximately 20 epochs. This observation indicates that the model's capacity for generalization on unseen data did not significantly improve beyond this point, despite continued training. Potential explanations for this plateau could include overfitting to the training set or reaching the inherent limitations of the model's architecture for our specific dataset characteristics. We closely monitored key metrics, such as precision, recall, and F1-score, during training and validation to evaluate model performance. Our detailed results, presented in Table 2, demonstrate that the Vision Transformer model achieved performance metrics that were closely comparable to those obtained using the VGG16 architecture, which served as a benchmark in our study. This suggests that, while the ViT model did capture complex patterns and relationships in our data, its overall effectiveness in our context was on par with a traditional convolutional neural network architecture. This finding raises interesting considerations regarding the choice of model architecture for our particular problem and the potential need for further optimization or adjustments to the training strategy for the ViT model.

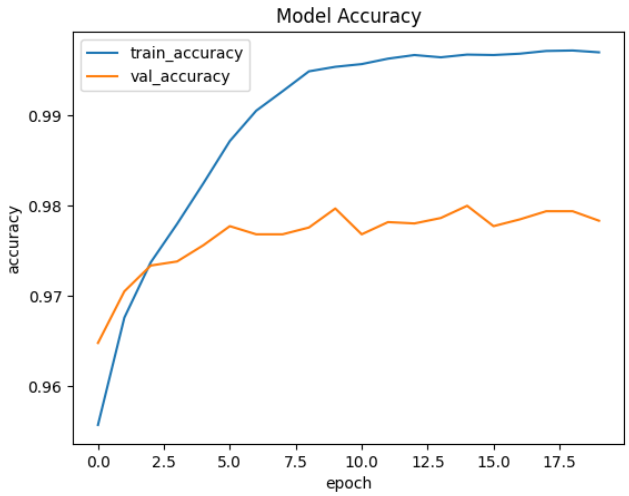

Fig. 3. Model Training for ViT (with 20 epochs)

3.2.3 LLM based Approach:

Approach 1: We used the LLAMA-Vision-11B model for our purpose. We created a detailed prompt using Deepseek-R1 to identify gambling related items in the image. We provided the vision model with the generated prompt and the image for classification. The LLM was prompted with an image as an input to output a structured JSON with a boolean output for gambling as 1 and not gambling as 0. We got 19.21% false positives (non-gambling classified as gambling) and 8.47% false negatives on the validation dataset. Based on the results of this approach, for our usecase, LLMs were performing poorly at discriminative / classification tasks compared to CNN based models or transformer models.

Approach 2: We used a sequence of steps to generate the propensities shown in Fig 7. Firstly, we prompted DeepSeek-R1 to generate a prompt that would help classify gambling images. We also manually inspected the images to gather attributes that would indicate a gambling image. We passed on this information to DeepSeek-R1 as well. This helped DeepSeek-R1 to give us a

further refined prompt, which would be used for the LLAMA-Vision-11B model. For the instruction prompt, which is used in the LLM2Vec technique described below, we prompted DeepSeek-R1 to generate a few gambling image examples along the lines of the manual image attribute information we collected. The instruction prompt, therefore, was a classification prompt with well-defined examples. We adopted LLM2Vec [7], a state-of-the-art unsupervised approach, to transform decoder-only language models into robust text encoders. By leveraging its parameter-efficient design and superior ability to generate contextualized text representations, LLM2Vec enables us to effectively model complex patterns in textual data, aligning with the nuanced requirements of our gambling application. We gave the image description and instruction prompt as an input to the LLM2Vec encoder. We then trained a logistic regression classifier, with saga solver and balanced class weights, on the encoded descriptions against the labels. This greatly improved the performance against Approach 1 and the earlier described models – VGG16 and ViT.

LLAMA Vision 11B Prompt:

Analyze this image and list all observable elements that could suggest it is related to risk-for-reward mechanics, monetary wagers, or casino-style activities. If absent, list elements that suggest generic advertising, entertainment, or other purposes.

Focus only on these categories: Monetary/Financial Elements: Casino chips, coins, banknotes, piles of cashCurrency symbols ($, €, ¥), free
spins" or "bonus" text. Prize amounts, jackpot displays, "win" or "wager" in text

Game Mechanics: Playing cards, dice, roulette wheels, slot machines. Poker tables, numbered balls (e.g., lottery), sports odds. Progress bars, "level up" indicators, timer countdowns

Ad/Creative Elements (if no gambling cues): Brand logos, app store badges, celebrity endorsements Neutral graphics (e.g., landscapes, abstract art) Call-to-action text ("Download Now," "Try Free")

Rules:
Only list items verbatim (e.g., "red poker chips," "$100 bill," "slot machine animation"). No summaries (avoid phrases like "this is gambling-related"). No assumptions—only include visible/textual elements.

Example Output:
Three golden casino chips stacked
Text: "Daily Bonus: 500 Coins" Green felt table with card symbols (:spades::hearts::diamonds::clubs:) Spin-the-wheel animation in background

Fig. 4. LLAMA Vision 11B Prompt

Instruction Prompt:

Classify as gambling or non-gambling. Some examples for gambling images:

Winning & Jackpot Focused:

"A euphoric gambler celebrating a massive slot machine jackpot win, coins pouring out, with a bright 'WINNER' sign flashing

and a disclaimer: 'Gambling involves risk. Play responsibly.'"

"A poker player revealing a royal flush, opponents groaning, with a pile of cash and chips pushed toward them—text overlay: 'Know when to walk away.'"

"A lottery ticket being scratched off, revealing a '$1,000,000' prize, with fine print at the bottom: 'Odds of winning vary. Bet with caution.'"

Risk & Luck-Based Imagery:

"A roulette wheel spinning, a nervous bettor watching the ball, with a casino sign in the background: 'Luck is a factor. Don't chase losses.'"

"A gambler hesitating before placing a high-stakes bet, a shadowy casino backdrop, with text: 'Risk is part of the game. Set limits.'"

"A slot machine screen displaying 'BONUS ROUND—YOU WON $500!' with a small disclaimer: 'Results are random. Play for fun, not profit.'"

"You've Already Won" or "Try Your Luck" Themes:

"A casino billboard glowing with 'TRY YOUR LUCK TODAY!' alongside images of dice, cards, and a tiny 'Terms and conditions apply.'"

"A smartphone screen showing a 'CONGRATS! YOU WON $200 CASH!' pop-up from a gambling app, with a 'Claim Prize' button and risk warning in small text."

"A gambler holding a golden ticket with 'INSTANT CASH PRIZE' stamped on it, surrounded by slot machines, and a subtle 'Gambling can be addictive.'"

"A blackjack table with a dealer smiling, pushing chips toward a player, text overlay: 'Feeling lucky? Remember—the house always has an edge.'"

Fig. 5. Instruction Prompt for LLM2Vec

The response for Fig. 6. clearly highlights the "gambling elements" within the image. It detects the "Rewards", "Gold Coins", "Fortune Wheel" which is at the bottom left part. As a result, the response is classified as Gambling.

Fig. 6. Sample Gambling Image

Response:

Based on the presence of the words ""Rewards"" and the wheel of fortune, the image appears to be advertising a gambling game for older aged people. The presence of gold coins at the bottom of the image adds to the suspicion that this is a gambling game for older players. The image is likely designed to entice mature players by describing the rewards they will receive after accumulating certain amounts of coins. The presence of a balloon and a gift box further reinforces this idea.

Ensemble Model

We combined the output propensities got from the VGG16, ViT and LLM2Vec model outputs and applied min-max normalization over the propensities and then trained a logistic regression classifier on them as shown in Fig 8. We observe that this further improves our metrics of interest.

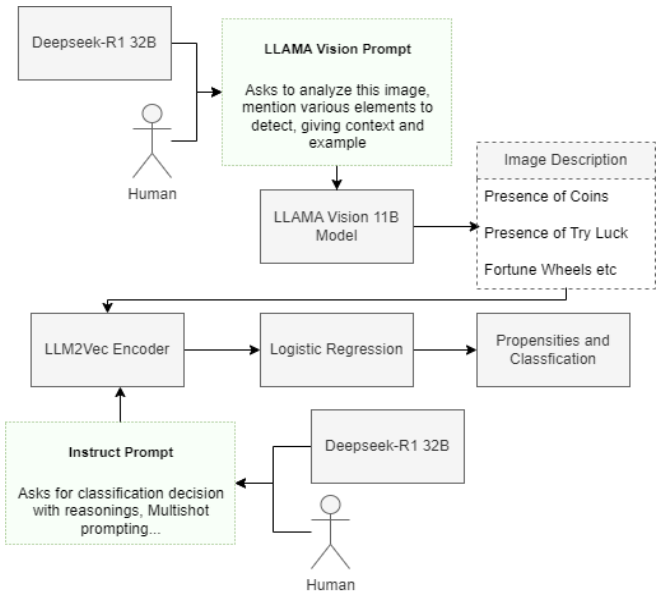

Fig. 7. Approach 2 For LLM2Vec based classification

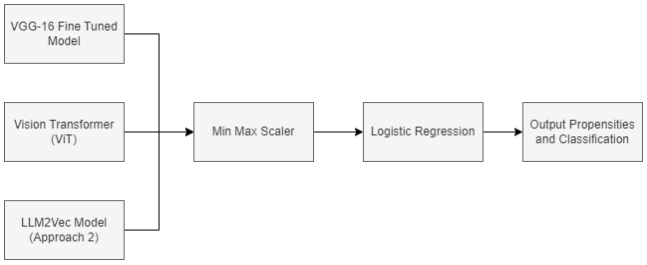

Fig. 8. Ensemble Model

4 RESULTS AND DISCUSSIONS

4.1 Evaluation Metrics:

We evaluated the model performance using the following metrics:

- FNR - Corresponds to actual gambling images classified as non-gambling.

$$FNR = \frac{FN}{FN + TP}$$

- FPR - Corresponds to actual non-gambling images classified as gambling.

$$FPR = \frac{FP}{FP + TN}$$

- F1-score - Evaluate across models

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.2 System and Hardware Specifications:

To conduct the various experiments involving LLMs, we have used the EC2 instance g5.12xlarge (4-GPUs with GPU architecture Nvidia A10g). AMI used was Deep Learning Base OSS Nvidia Driver GPU AMI (Ubuntu 22.04).

4.3 Results:

The Table 1 shows the relative performance of the models. We see that the ensemble provided the best performance followed by the LLM2Vec model approach with respect to the gambling image classification problem. We applied two propensity thresholds to break the prediction into three groups as shown in Table2. The manual review team need not spend time on 91% of images predicted as "Not gambling" and 4% of the images predicted as "Gambling". Our model, therefore, **saves 95% of the time and effort** with only **5.36%** of gambling images (i.e. 17 gambling misclassified out of (17+59+241=317)) and **0.61%** of non-gambling images misclassified.

5 CONCLUSIONS AND FUTURE WORK

This study demonstrates the potential of using LLMs and ensemble techniques to enhance the efficiency of the creative review process for detecting gambling advertisement images. We plan to implement models like CLIP to improve performance. We also observe that the models we implemented are not able to use financial disclaimers (generally present at the end of an image) when making the classification decision. We plan to use OCR based models to help classification in such cases.

Model	Dataset	FPR	FNR	F1-Score
VGG16	Test	2.32	12.16	0.8208
	Val	2.06	12.31	0.8388
ViT	Test	3.22	12.72	0.7342
	Val	3.55	12.56	0.7148
LLM2Vec	Test	1.98	11.68	0.8516
(Approach2)	Val	1.83	11.91	0.8414
Ensemble	Test	1.59	11.02	0.8806
	Val	1.57	10.97	0.8923

Table 1: Results from different models

Predicted Group	Actual 0	Actual 1	%age of miss	%age of total
Not Gambling	6061	17	0.61%	91%
Manual check	234	59	-	4%
Gambling	39	241	5.36%	4%

Table 2: Final Results of Validation Set with Manual Review

6 ACKNOWLEDGMENT

All the work done was part of Samsung Ads. We are thankful to Samsung Ads for providing the necessary data and resources to conduct our research and experiments.

REFERENCES

[1] S. Thomas, M. C. I. van Schalkwyk, M. Daube, H. Pitt, D. McGee and M. McKee, "Protecting children and young people from contemporary marketing for gambling," Health Promotion International, vol. 38, March, 2023. DOI: https://doi.org/10.1093/heapro/daac194

[2] Li L, Gou G, Xiong G, Cao Z, Li Z. Identifying gambling and porn websites with image recognition. InAdvances in Multimedia Information Processing–PCM 2017: 18th Pacific-Rim Conference on Multimedia, Harbin, China, September 28-29, 2017, Revised Selected Papers, Part II 18 2018 (pp. 488-497). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-77383-4_48

[3] Chen, Y., Zheng, R., Zhou, A., Liao, S. and Liu, L. 2020. Automatic Detection of Pornographic and Gambling Websites Based on Visual and Textual Content Using a Decision Mechanism. Sensors. 20, 14 (Jul. 2020), 3989. DOI: https://doi.org/10.3390/s20143989.

[4] Jichen Zhang, Qiang Duan, Enhao Zhan, Zizhong Wei, Changsheng Liu, Kai Jiang, and Rui Li. 2024. Zero-Shot Harmful Image Recognition Based On Innovative Dataset Construction and CLIP Embedding. In Proceedings of the 2023 International Conference on Information Education and Artificial Intelligence (ICIEAI '23). Association for Computing Machinery, New York, NY, USA, 328–333. DOI: https://doi.org/10.1145/3660043.3660102

[5] Simonyan, K. and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv DOI: https://doi.org/10.48550/arXiv.1409.1556

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. DOI: https://doi.org/10.48550/arXiv.2010

[7] BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N. and Reddy, S. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. arXiv DOI: https://doi.org/10.48550/arXiv.2404.05961