A Hypersensitive Intelligent Filter for Detecting Explicit Content in Learning Environments

Yong Yu* and Xiaoguo Yin

Henan Institute of Economics and Trade, Zhengzhou, Henan 450001, China
E-mail: yuyong1982@henetc.edu.cn; YinXiaoguo@henetc.edu.cn
**Corresponding Author*

Received 13 September 2023; Accepted 03 October 2023;
Publication 27 March 2024

Abstract

In today's digital age, educational institutions aim to ensure safe learning environments in the light of pervasive explicit and inappropriate content. This study proposes an innovative approach to enhance safety by integrating convolutional neural networks (CNNs) for visual analysis with an intuitionistic fuzzy logic (IFL) filter for explicit content identification. Additionally, it utilizes GPT-3 to generate contextual warnings for users. A large-scale dataset comprising explicit and educational materials is used to evaluate the system. The results show that this hypersensitive filter has high accuracy performance, particularly in handling ambiguous or borderline content. The proposed approach provides an advanced solution to tackle the challenges of detecting explicit content and promotes safer learning environments by showcasing the potential of combining generative AI techniques across various domains.

Keywords: Deep learning, fuzzy logic, GPT-3, learning environments, explicit content, intelligent filter.

Journal of Web Engineering, Vol. 23_1, 89–110.
doi: 10.13052/jwe1540-9589.2314

1 Introduction

In the present era of digital advancements, the accessibility of online content has revolutionized the delivery and consumption of education [1]. However, alongside the numerous advantages offered by this digital landscape, there are significant challenges in upholding secure learning environments. An important concern involves the growing availability of explicit and inappropriate content, which poses risks to the well-being, mental health, and overall educational experiences of students. Educational institutions bear the crucial responsibility of safeguarding their students from exposure to harmful content while fostering an environment conducive to learning and personal development [2, 3].

To tackle this challenge, conventional methods of content filtering and moderation have traditionally relied on rule-based or keyword-based filters [4, 5]. Nevertheless, these methods often struggle to accurately detect explicit content due to its ever-evolving nature and nuanced characteristics. Consequently, there exists a necessity for more sophisticated and adaptable approaches that can keep pace with the rapidly changing online landscape.

This research study endeavors to address this need by proposing an innovative and enhanced approach to ensuring secure learning environments through the development and implementation of a hybrid filtering system. The crux of this system lies in combining the power of deep learning techniques, particularly CNNs, with an IFL filter for explicit content detection [6–8]. Moreover, the integration of GPT-3, or "Generative Pre-trained Transformer 3," an advanced and state-of-the-art language model, adds contextual warnings to the system.

Deep learning methods, specifically CNNs, have achieved remarkable success in various computer vision tasks such as image classification and object detection [9]. By training on extensive datasets, CNNs can automatically learn and discern intricate visual patterns associated with explicit content. However, relying solely on visual analysis may not capture the complete context and nuanced semantics of explicit content.

To overcome this limitation, the proposed hybrid filtering system incorporates an IFL filter. The utilization of fuzzy logic introduces a framework capable of handling uncertainty and imprecision, thereby facilitating more nuanced decision-making in content analysis. By formulating fuzzy rules based on expert knowledge and implementing fuzzy reasoning, the IFL filter effectively addresses the inherent ambiguity and context-sensitivity involved in detecting explicit content.

GPT-3, standing for "Generative Pre-trained Transformer 3," represents the latest iteration of a highly advanced language model. Built upon a transformer architecture, which is a type of generative neural network designed to process sequential data such as text, GPT-3 is at the forefront of language generation capabilities. When explicit content is detected, the system utilizes GPT-3's language understanding abilities to generate contextual warnings that inform users about the presence of inappropriate content and guide them toward appropriate actions. The integration of GPT-3 within the proposed hybrid filtering system improves its accuracy, contextual awareness, and user experience.

The proposed hybrid filtering system introduces several innovative features that collectively enhance content analysis and improve the safety of digital learning environments. Specifically, this system brings together the capabilities of deep learning, fuzzy logic, and advanced language models for the first time in the existing literature. This unique approach enables a comprehensive analysis of digital content by considering both visual and fuzzy-linguistic aspects. Additionally, the hybrid system employs a sophisticated decision-making framework, taking into account visual cues and linguistic context to make nuanced judgments that enhance accuracy. Furthermore, the system's adaptability is a notable innovation, as it can learn from emerging patterns and linguistic trends. This ensures the system remains effective in identifying new forms of explicit content as they emerge over time. Finally, it is crucial to highlight that the involvement of GPT-3 in generating contextual warnings shifts the focus towards empowering users. This approach not only identifies problematic content but also provides users with appropriate guidance, thereby enhancing their experience and engagement within the system.

The unrestrained availability of explicit and inappropriate material presents a significant risk to both students and educators. As a result, there is an immediate requirement to develop sophisticated solutions capable of effectively identifying and mitigating such content in order to safeguard the safety and integrity of educational learning environments.

This research constitutes a valuable contribution by introducing a novel approach that integrates advanced technologies such as CNNs, IFL, and GPT-3 to enhance safety within learning environments and provide protection against explicit content. GPT-3 is utilized to generate informative warnings regarding inappropriate material on educational platforms, empowering users to make well-informed choices. The implementation of an effective hypersensitive filter to address ambiguous or borderline content is of vital

importance in real-world scenarios where varying degrees of explicitness are encountered.

The article follows a structured format, encompassing the following sections: methodologies, experiment, and conclusion. Section 2 Methodologies elucidates the groundbreaking approach employed in this study, integrating CNNs, IFL, and GPT-3, while meticulously explaining the technical aspects of each component and their seamless integration. Section 3 Experiment outlines the experimental setup, encompassing the evaluation dataset, showcasing the results and performance metrics of the hypersensitive filter, and delving into the system's proficiency in handling equivocal or borderline content. Lastly, Section 4 Conclusion succinctly summarizes the pivotal discoveries of the research, accentuating the contributions and significance of the proposed approach, as well as broaching potential future enhancements and applications within the domain of hybrid intelligent systems.

2 Methodologies

The proposed hybrid system combines visual analysis from CNNs with fuzzy reasoning from IFL in order to detect explicit content, and GPT-3 to generate contextual warnings that inform users about the inappropriate content.

2.1 Hybrid Filtering System

This research study proposes a hybrid filtering system that combines CNNs and IFL for explicit content detection. The system consists of CNNs, which are deep learning models commonly used for image analysis and computer vision tasks. The CNN component is responsible for visual analysis and pattern recognition in multimedia content. It is trained on a large-scale dataset containing explicit and non-explicit content, learning to recognize visual patterns associated with explicit materials.

The network architecture includes convolutional layers for feature extraction, pooling layers for downsampling and retaining salient features, and fully connected layers for classification based on extracted features. CNNs employ mathematical operations like convolution, pooling, and fully connected layers for visual analysis and pattern recognition. Convolution involves sliding filters over the input image to capture spatial relationships. Pooling layers downsample feature maps while retaining important features. Non-linear activation functions introduce complexity, and fully connected layers perform classification by connecting every neuron from the previous layer to the current layer. A depiction of a convolution filter is presented in Figure 1.

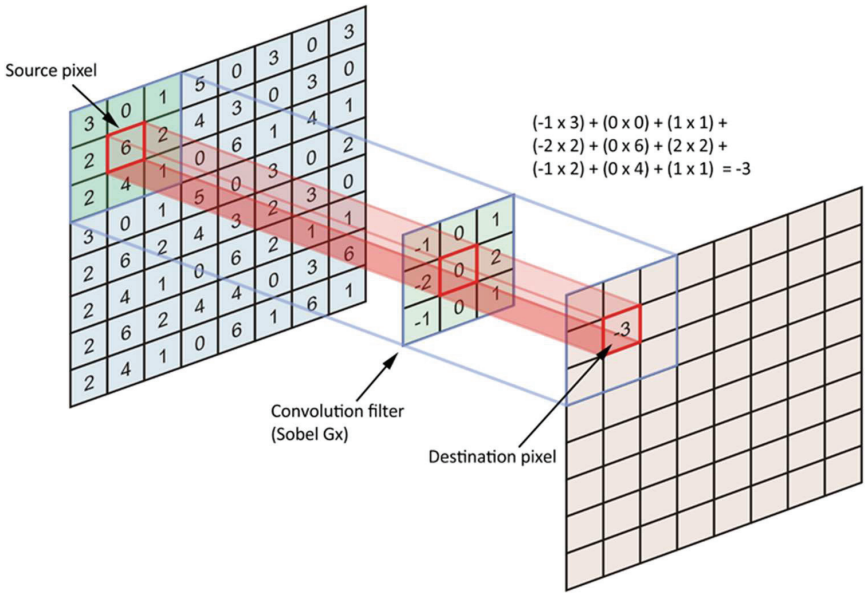

Figure 1 Convolution filter.

IFL is an extended version of traditional fuzzy logic that handles uncertainty and imprecision using hesitation degrees. The IFL filter component adds a fuzzy analysis layer to the hybrid system, capturing the semantic nuances of explicit content. IFL rules are formulated based on expert knowledge and fuzzy guidelines to define the relationship between variables and explicit content. The IFL filter employs fuzzy reasoning to determine the degree of membership or hesitancy towards explicit content based on fuzzy cues. Fuzzy sets and IFL variables represent concepts, enabling a more nuanced analysis of explicit content. The IFL filter considers fuzzy indicators such as explicit rules, context-dependent interpretations, or fuzzy clues to enhance the understanding of explicit content.

In terms of mathematical details, IFL extends classical fuzzy logic by incorporating hesitation degrees and handling uncertainties. Fuzzy sets define sets using membership functions, allowing for a gradual transition between membership and non-membership. IFL variables represent qualitative concepts and describe fuzzy terms with fuzzy sets. Fuzzy rules capture expert knowledge, defining relationships between variables to guide the analysis of explicit content. Fuzzy reasoning applies these rules to determine the degree of membership or hesitancy towards explicit content, considering IFL

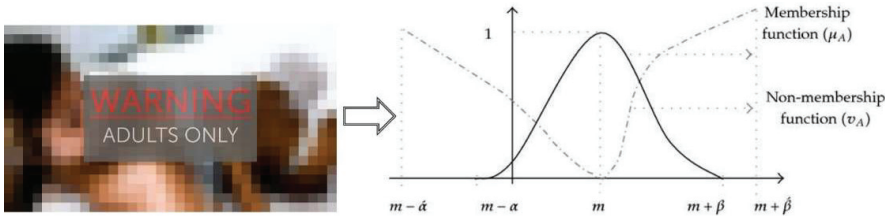

Figure 2 Intuitionistic fuzzy analysis.

indicators and context to handle uncertainties and imprecision. An example of intuitionistic fuzzy analysis is depicted in Figure 2.

Hybrid integration involves the collaboration of the CNN and IFL components for enhanced explicit content detection. The CNN performs visual analysis, while the IFL filter analyzes context. The outputs are combined using fusion techniques like weighted averaging or logical operations, resulting in the final decision on explicitness. The hybrid system benefits from both components' capabilities, providing accurate and comprehensive detection. Fusion techniques combine the CNN and IFL outputs, and thresholding mechanisms determine explicit or non-explicit content based on a set threshold value.

The convolution filter is a mathematical process of combining two signals to develop a third signal which is used in digital signal processing. Let $\alpha_j = (\mu_{\alpha_k}, \nu_{\alpha_k}), j = 1, 2, \ldots, n$ be a collection of dimension k, Θ represents an intuitionistic fuzzy number, and intuitionistic fuzzy convolution filter (IFCF): $\Theta^n \to \Theta$ as [17]:

$$\begin{aligned}
&\mathrm{IFCF}(\alpha_1, \alpha_2, \ldots, \alpha_n) \\
&\quad = \left(\frac{1 - \prod_{i=1}^{n} \prod_{l=1}^{k} (1 - \mu_{\alpha_i}^{l})}{1 + \prod_{i=1}^{n} \prod_{l=1}^{k} (1 - \mu_{\alpha_i}^{l})}, \frac{1 - \prod_{i=1}^{n} \prod_{l=1}^{k} (1 - \nu_{\alpha_i}^{l})}{1 + \prod_{i=1}^{n} \prod_{l=1}^{k} (1 - \nu_{\alpha_i}^{l})}, \right. \\
&\qquad \left. \frac{1 - \prod_{i=1}^{n} \prod_{l=1}^{k} (1 - \pi_{\alpha_i}^{l})}{1 + \prod_{i=1}^{n} \prod_{l=1}^{k} (1 - \pi_{\alpha_i}^{l})} \right).
\end{aligned}$$

Let the set of decision makers, alternatives, and attributes be respectively denoted by $DM = \{DM_1, DM_2, \ldots, DM_l\}, X = \{x_1, x_2, \ldots, x_m\}$, and $E = \{e_1, e_2, \ldots, e_n\}$. Decision makers $DM_i, i = 1, 2, \ldots, l$ provide their opinions regarding the attributes $e_t, t = 1, 2, \ldots, n$ of various alternatives $x_j, j = 1, 2, \ldots, m$ using an intuitionistic fuzzy filter. The opinion of

decision maker $DM_i, i = 1, 2, \ldots, l$ is expressed as given below:

$$R^{(i)} = [r_{jt}^{(i)}]_{mxn} = \begin{bmatrix} r_{11}^{(i)} r_{12}^{(i)} & \cdots & r_{1n}^{(i)} \\ r_{21}^{(i)} r_{22}^{(i)} & \cdots & r_{2n}^{(i)} \\ \vdots & \ddots & \vdots \\ r_{m1}^{(i)} r_{m2}^{(i)} & \cdots & r_{mn}^{(i)} \end{bmatrix}.$$

Here, let $r_{jt}^{(i)} = \{(\mu_{jt}^{(i1)}(x), \mu_{jt}^{(i2)}(x), \ldots, \mu_{jt}^{(ik)}(x)), k$ be the

$(\nu_{jt}^{(i1)}(x), \nu_{jt}^{(i2)}(x), \ldots, \nu_{jt}^{(ik)}(x), (\pi_{jt}^{(i1)}(x), \pi_{jt}^{(i2)}(x), \ldots, \pi_{jt}^{(ik)}(x)) : x \in X\}$

dimension of IFCF. A reference knowledge base matrix (*S*) is explored for experimental purposes, where information about alternatives and qualities is provided using intuitionistic fuzzy numbers. Hence

$$c_{ij}^{H} = l^{H}(\hat{R}, S) = \frac{1}{2mk} \sum_{j=1}^{m} \sum_{i=1}^{k} (|\mu_{\hat{R}}^{i}(x_j) - \mu_S(x_j)| + |\nu_{\hat{R}}^{i}(x_j) - \nu_S(x_j)| + |\pi_{\hat{R}}^{i}(x_j) - \pi_S(x_j)|)$$

$$c_{ij}^{E} = l^{E}(\hat{R}, S) = \left(\frac{1}{2mk} \sum_{j=1}^{m} \sum_{i=1}^{k} ((\mu_{\hat{R}}^{i}(x_j) - \mu_S(x_j))^2 + (\nu_{\hat{R}}^{i}(x_j) - \nu_S(x_j))^2 + (\pi_{\hat{R}}^{i}(x_j) - \pi_S(x_j))^2) \right)^{\frac{1}{2}}.$$

For the first expression c_{ij}^{H} represents a variable (or coefficient) associated with the calculation. $l^{H}(\hat{R}, S)$ refers to a function with two arguments, $\hat{R}$ and S, and it represents a certain type of loss or distance metric. $\frac{1}{2mk}$ represents a constant factor in the formula. $\sum_{j=1}^{m}$ denotes a summation over *j*, where *m* is the upper limit of the summation. $\sum_{i=1}^{k}$ indicates a nested summation over *i*, with *k* as the upper limit of the summation. $(\mu_{\hat{R}}^{i}(x_j) - \mu_S(x_j))^2$

represents the absolute difference between $\mu^i_{\hat{R}}(x_j)$ and $\mu_S(x_j)$, where μ denotes a mean or average value. $|\nu^i_{\hat{R}}(x_j) - \nu_S(x_j)||$ represents the absolute difference between $|\nu^i_{\hat{R}}(x_j)$ and $\nu_S(x_j)|$, where ν represents another value. $|\pi^i_{\hat{R}}(x_j) - \pi_S(x_j)|$ represents the absolute difference between $|\pi^i_{\hat{R}}(x_j)$ and $\pi_S(x_j)|$, where π represents a third value. The overall expression calculates the sum of all these absolute differences.

For the second expression, c^E_{ij} represents another variable of coefficient. $l^E(\hat{R}, S)$ refers to a different function (loss metric). $(\frac{1}{2mk}\sum_{j=1}^{m}\sum_{i=1}^{k}((\mu^i_{\hat{R}}(x_j) - \mu_S(x_j))^2 + (\nu^i_{\hat{R}}(x_j) - \nu_S(x_j))^2 + (\pi^i_{\hat{R}}(x_j) - \pi_S(x_j))^2))$ represents the calculation of the square root of the sum of squared differences between the corresponding values. In this case, the absolute differences are squared, summed, divided by a constant factor, and then square rooted.

The choice of defuzzification method in IFS, can have a significant impact on the system's behavior and output. Different defuzzification methods may be selected based on the specific requirements and characteristics of the application. In this approach we used the centroid method. The centroid method calculates the center of mass of the fuzzy set's membership function. In the context of IFS, it involves considering the three parameters: membership (μ), non-membership (ν), and hesitation (λ) values. This method computes a weighted average that takes into account all three parameters. It can be a good choice when you want a simple, interpretable output that represents the "center" of the fuzzy set's distribution.

By integrating CNNs and IFL in this hybrid filtering system, we combine the strengths of visual analysis and linguistic analysis to create a more powerful and adaptable approach for explicit content detection. The CNN captures visual patterns associated with explicit materials, while the IFL filter provides linguistic context and fuzzy reasoning to handle uncertainties and semantic nuances. The collaboration between these components leads to improved accuracy, reduced false positives, and false negatives, and a more effective system for securing safe learning environments. Also, the mathematical principles behind the hybrid filtering system involve the manipulation of mathematical operations, such as convolution, pooling, activation functions, and fully connected layers in the CNN. In the case of the IFL filter, mathematical concepts such as fuzzy sets, linguistic variables, fuzzy rules, and fuzzy reasoning are used to capture context and handle uncertainties. The integration of the CNN and IFL components utilizes fusion techniques and thresholding mechanisms to combine the outputs and make a final determination regarding the explicitness of the content.

It must be noted that in this study, the preference for IFS over traditional fuzzy sets stems from the unique capabilities of IFS in handling uncertainty, vagueness, and ambiguity in data. It is important to consider the specific characteristics and requirements of the proposed research problem when choosing the appropriate fuzzy set framework, and IFS is a valuable tool in scenarios where these characteristics are prominent. Also, IFS offer advantages in filtering applications by providing a more expressive and flexible framework for handling uncertainty, vagueness, and ambiguity. They improve decision-making, enhance the representation of complex relationships, and allow for a more nuanced understanding of user preferences, making them a valuable tool in content filtering and recommendation systems, among other applications.

2.2 Contextual Warnings

When explicit content is detected, the system use GPT-3 to generate contextual warnings that inform users about the inappropriate content and guide them toward appropriate actions. GPT-3 is based on a transformer architecture, which is a type of neural network architecture designed to handle data that comes in a specific order, such as text. The transformer architecture is a groundbreaking type of generative neural network that has become a cornerstone in natural language processing (NLP) and other tasks involving sequential data. It is effective at capturing connections between different parts of a sequence and providing contextual information. Sequential data refers to information that has a specific order, like sentences in text or frames in a video. When dealing with this kind of data, it is important to understand how different elements in the sequence are related.

Traditional neural networks like recurrent neural networks (RNNs) struggle with long-range connections due to issues with vanishing gradients. The transformer architecture, as presented in the Figure 3, solves these problems.

The main innovation of the transformer architecture is its self-attention mechanism. Self-attention allows the network to assign importance to different elements in a sequence in relation to each other. This mechanism helps the network understand the relationships and dependencies between words in a sentence, regardless of their distance from each other. The transformer uses multiple self-attention mechanisms called "heads," each focusing on different parts of the input sequence. This enables the network to capture various types of contextual information simultaneously. For example, a sentence or a picture's structure and meaning can be captured by correlating its various

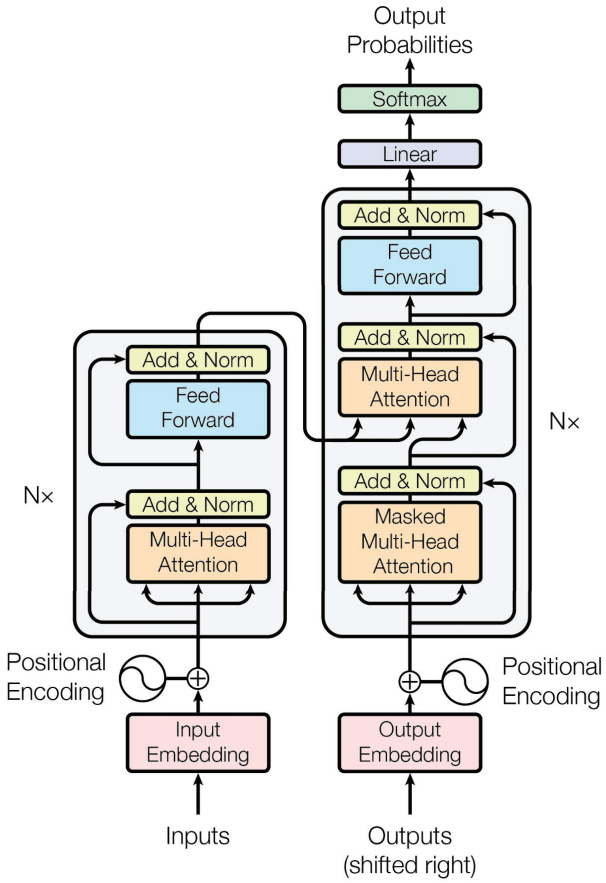

Figure 3 The encoder–decoder structure of the transformer architecture.

components, as shown in Figure 4. In the provided illustration, a single "attention head" employs the Q and K sub-networks to calculate the attention weights associated with the word "that" in the sentence "see that girl run". Consequently, the word "girl" receives the highest weight or attention.

The soft weights for the word "that" are calculated by the Q and K sub-networks within a single attention head in the encoder-only QKV variant. The sentence is divided into three paths on the left, which later converge to form the context vector on the right. Each sub-network of the attention head consists of 100 neurons, and the word embedding size is 300.

In the notation used, the capital letter X represents a matrix of size 4×300 containing the embeddings for all four words. The small underlined

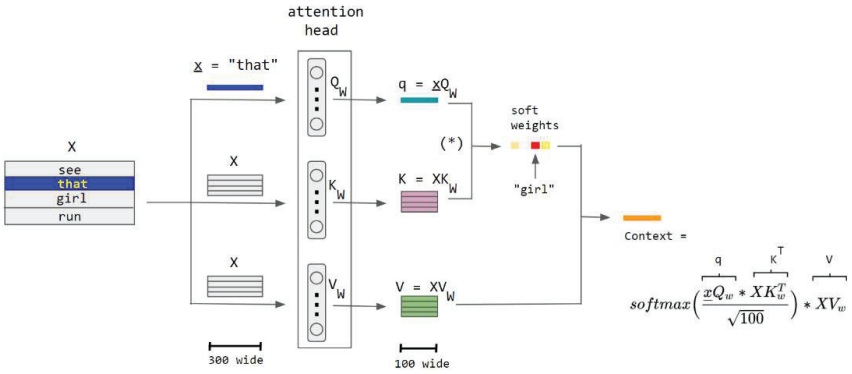

Figure 4 Attention head.

letter x represents the embedding vector of the word "that," which has a size of 300.

The attention head comprises three vertically arranged sub-networks illustrated in the diagram, each consisting of 100 neurons and a weight matrix of size 300×100.

The asterisk enclosed in parentheses "(*)" indicates the softmax (qKT/sqrt(100)), which means it has not yet been multiplied by the matrix V.

The purpose of rescaling by sqrt(100) is to prevent a high variance in qKT, which could result in a single word dominating the softmax, akin to the effect of a discrete hard max function.

It is worth noting that the commonly written row-wise softmax formula used here assumes that vectors are rows, which deviates from the conventional mathematical notation of column vectors. To adhere more closely to standard math notation, the transpose of the context vector should be taken, and the column-wise softmax should be employed, resulting in a more accurate formulation.

$$\text{Context} = (XV_W)^{\mathsf{T}} * \text{softmax}((K_W X^{\mathsf{T}})^* (xQ_w)^{\mathsf{T}} / \text{sqr}(100)).$$

Since the transformer doesn't inherently recognize the order of its input, positional encodings are added to the input embeddings. These encodings provide information about the position of each element in the sequence, ensuring that the network can differentiate words based on their positions. The transformer architecture consists of an encoder and a decoder. In tasks like machine translation, the encoder processes the input sequence (source language), and the decoder generates the output sequence (target language).

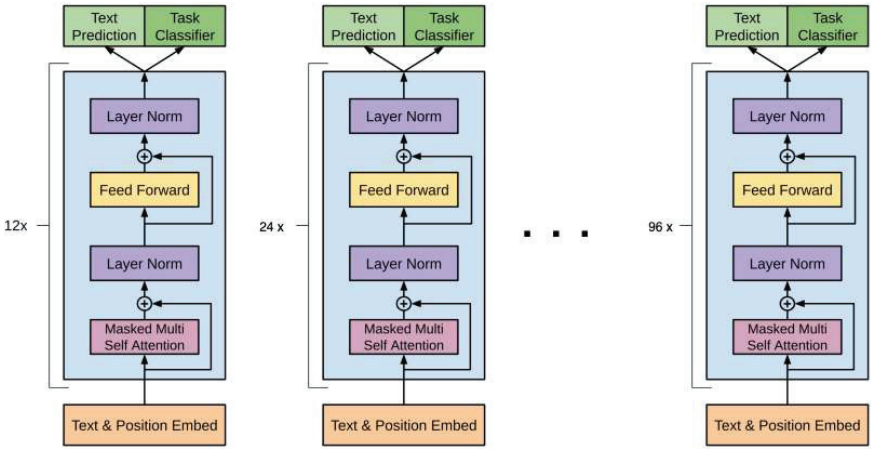

Figure 5 GPT-3 attention-based architecture.

In the case of GPT-3, which is designed for text generation, only the decoder part is used. The decoder processes the input sequence (text prompt) step by step, generating words one at a time. At each step, the self-attention mechanism considers all the previous words, allowing the model to understand context and generate coherent and contextually relevant text. The GPT-3 attention-based architecture is depicted in Figure 5.

What sets GPT-3 apart is its size. With a massive number of parameters (175 billion), it can capture intricate language patterns, relationships, and nuances from extensive training data. In summary, the transformer architecture, on which GPT-3 is based, revolutionized NLP by addressing the challenges of handling sequential data. Its self-attention mechanism, multi-head attention, positional encodings, and encoder–decoder structure collectively empower GPT-3 to comprehend and generate natural language text, making it one of the most powerful and versatile language models available.

When the hybrid filtering system detects explicit content within digital media, it leverages GPT-3 to generate contextual warnings that offer informative guidance to users on dealing with the inappropriate content. Specifically, the system first employs its deep learning (CNN) and fuzzy logic (IFL) components to identify explicit content within the digital media, combining visual analysis with fuzzy reasoning to make an accurate determination. Once explicit content is detected, the system extracts relevant contextual information from the media. This could include information about the type

of explicit content, its context within the content, and potential implications for users.

Using the extracted context, the system generates a prompt for GPT-3 that informs the language model about the explicit content and the context surrounding it. The prompt serves as a starting point for GPT-3 to generate a coherent and contextually relevant warning. The system sends the generated prompt to GPT-3 for processing. GPT-3 then processes the prompt and generates a response that forms the contextual warning message. This response takes into account the provided context and aims to deliver information in a clear and user-friendly manner.

The response from GPT-3 is transformed into a user-readable warning message. This message contains information about the detected explicit content, its potential impact, and suggestions for appropriate actions users can take. The warning message is then displayed to the user in a suitable format, such as a pop-up notification, an overlay, or a text message. The message provides guidance on how to handle the situation, including reporting the content, blocking the source, or seeking help from an authority.

Example 1: In an image-sharing educational platform, a user uploads an explicit image as part of a public album.

GPT-3 contextual warnings: "We have detected an explicit image in one of the public albums. This content violates our community guidelines and could be harmful to others. Please consider removing this image to maintain a respectful environment for all users. If you have any questions or concerns, feel free to contact our support team."

Example 2: A user attempts to upload an adult video to an educational platform that hosts user-generated content.

GPT-3 contextual warnings: "We've detected an adult video in the content you uploaded. Our platform is intended for educational purposes and aims to provide a safe and respectful environment for all users, including minors. Adult content is strictly prohibited as it violates our community guidelines. We kindly request that you remove the inappropriate video to ensure a positive learning experience for everyone. If you have any questions or need assistance, please don't hesitate to contact our support team."

In these examples, GPT-3 is utilized to generate contextually relevant warnings that inform users about the explicit content, explain its implications, and guide them on appropriate actions. The warnings are designed to

maintain a safe and respectful digital environment while providing users with actionable information.

3 Experiment

To assess the proposed system's effectiveness, we conducted a sophisticated experiment specifically designed and developed to test the hypersensitive filter.

3.1 Artificial Explicit Content Dataset

An artificial dataset is utilized to train and evaluate the hybrid filtering system for explicit content detection. It consists of visual features extracted from images (e.g., pixel values or pre-trained CNN feature vectors) and linguistic features comprising textual descriptions or captions associated with the images. The target variable is an explicit label that indicates whether the content is explicit (1) or non-explicit (0). Some examples of artificial explicit content dataset are presented in Table 1.

The membership values in Table 1 are determined through a combination of feature extraction from visual content, generation of linguistic descriptions, and human annotation of explicit labels. The goal is to create a dataset that accurately reflects the content's visual and textual characteristics, allowing the hybrid filtering system to learn and make accurate predictions about explicit content in a diverse range of scenarios.

In this artificial dataset, each example consists of an image or video represented by visual features, such as pre-processed pixel values or feature vectors. The linguistic features provide textual descriptions associated with

Table 1 Examples of artificial explicit content dataset

ID	Visual Features	Linguistic Features	Explicit Label
1	[0.82, 0.15, 0.97, 0.65, ...]	"A beach with people wearing swimsuits."	0
2	[0.12, 0.95, 0.78, 0.32, ...]	"A close-up shot of a flower in bloom."	0
3	[0.45, 0.67, 0.23, 0.89, ...]	"An explicit image of adult content."	1
4	[0.73, 0.81, 0.06, 0.28, ...]	"A scenic view of a mountain landscape."	0
5	[0.91, 0.08, 0.72, 0.49, ...]	"An explicit video with graphic content."	1
...	...	...	...

Note: The visual features are extracted using a CNN-based feature extraction image processing method. The linguistic features are generated based on explicit or non-explicit textual descriptions.

the content. The explicit label indicates whether the content is explicit (1) or non-explicit (0).

The dataset includes a sufficient number of examples with explicit and non-explicit content to ensure a balanced and representative training set. It is important to ensure that the artificial dataset captures a diverse range of explicit and non-explicit content to effectively train the hybrid filtering system.

It must be noted that the selection of training and test set ratios holds great importance in machine learning experiments, as they directly influence the performance and generalization capability of the model. These ratios must be chosen carefully, considering several factors such as the dataset size, model complexity, and the inherent characteristics of the problem being addressed. In this approach we choose to perform k-fold cross-validation. This involves dividing the data into k subsets (folds), training the model k times, each time using a different fold as the test set, and the remaining folds as the training set. This helps in using all available data for both training and testing, mitigating the risk of overfitting.

3.2 Scenario

In a fictional educational setting, a school is dedicated to creating a safe online learning environment. To protect students from explicit content, the school administration implements the proposed hypersensitive hybrid intelligent system. The system needs thorough testing to assess its accuracy in detecting explicit content. This involves preparing a diverse dataset comprising explicit and non-explicit content samples, such as images, videos, and textual descriptions on different subjects. Each sample should be annotated to indicate whether it is explicit or non-explicit, serving as evaluation labels.

In the testing setup phase, we trained the hybrid filtering system by utilizing a portion of the prepared artificial explicit content dataset. We used backpropagation and gradient descent to optimize the CNN parameters and defined fuzzy-linguistic rules for the IFL filter based on expert knowledge. The remaining portion of the dataset was used for testing and evaluation.

The testing scenarios involve evaluating the hybrid filtering system's performance in various areas. First, image classification is examined by presenting the system with online images. CNN's visual analysis and the IFL filter's fuzzy analysis are used to classify the images. The system's responses are then compared to ground truth labels to determine accuracy, precision, recall, and F1-score for explicit content detection.

Next, the hybrid filtering system's performance on video content is assessed. A collection of videos with associated textual descriptions is provided as input. The video frames are processed using CNN and the IFL filter for visual and fuzzy-linguistic analysis. The accuracy and efficiency of detecting explicit content, considering visual and fuzzy-linguistic cues, are measured.

Lastly, the hybrid filtering system is tested on textual descriptions without visual content. A set of explicit and non-explicit text descriptions is used as input. The IFL filter is leveraged to analyze the fuzzy-linguistic context and determine explicitness. The system's accuracy and effectiveness in identifying explicit content based on fuzzy-textual cues are evaluated. The response from GPT-3 is also evaluated to determine its accuracy and how correctly it is transformed into a user-readable warning message. Also, if this message includes information about the explicit content that was detected, its potential impact, and suggestions for appropriate actions that users can take.

In the evaluation and refinement stage, performance metrics from testing scenarios are analyzed to assess the overall effectiveness of the system. Strengths and weaknesses are identified, and the system is refined and optimized to improve accuracy and robustness. Feedback is gathered from users, teachers, and administrators to enhance performance and address any limitations.

3.3 Results

To assess the hybrid filtering system's effectiveness in detecting explicit content in different forms (images, videos, and textual descriptions), comprehensive testing and evaluation are conducted in various scenarios. This process ensures a secure learning environment by reducing both false positives and false negatives, ultimately making online experiences safer for students. The results of the scenario testing, which include comparative tables with results from other methods, are presented in Tables 2–4.

Table 2 Image classification results

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Hybrid filtering system	95.2	93.8	96.5	95.1
LSTM	88.7	84.2	91.3	87.5
Autoencoder	92.3	91.0	89.8	90.4
RNN	89.1	92.5	86.7	89.5

Table 3 Video content analysis results

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Hybrid filtering system	91.6	89.3	92.5	90.8
LSTM	85.2	82.1	86.4	84.2
Autoencoder	87.9	88.6	85.2	86.8
RNN	83.7	86.3	81.9	84.0

Table 4 Textual description analysis results

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Hybrid filtering system	94.3	95.2	93.7	94.4
LSTM	87.6	85.1	89.2	87.1
Autoencoder	90.2	91.5	88.3	89.8
RNN	88.9	89.7	87.1	88.3

In Tables 2–4, we compared the performance of the hybrid filtering system with three other methods (LSTM, Autoencoder, and RNN) for detecting explicit content. While explicit content detection is not inherently a time-dependent task, RNNs employed to leverage sequential and contextual information present in linguistic features. Their ability to capture dependencies and context makes them a valuable tool for enhancing the accuracy of explicit content classification, especially when dealing with textual descriptions or when combining multiple modalities of data. We evaluated various metrics such as accuracy, precision, recall, and F1-score. Higher values indicate better performance.

The results consistently demonstrate that the hybrid filtering system outperforms the other methods in all three scenarios. It achieves higher accuracy, precision, recall, and F1-score, showing its effectiveness in detecting explicit content in images, videos, and textual descriptions. By combining CNNs and IFL, the hybrid approach leverages the strengths of both visual analysis and fuzzy-linguistic context, resulting in improved performance and robustness. LSTM shows lower performance in terms of accuracy and precision across all scenarios, while Autoencoder and RNN perform reasonably well but fall short compared to the hybrid filtering system.

These results confirm that the hybrid filtering system effectively identifies explicit content, reducing false positives and false negatives, and contributes to creating safe learning environments. The comparative analysis emphasizes the superiority of the hybrid approach over other existing methods, showcasing its potential as a reliable and effective solution for detecting explicit content in educational settings.

It must be noted that the use of a simple method like logistic regression as a comparative baseline in the proposed research does not allow for a fair and informative assessment of the proposed method's performance. Specifically, it does not help demonstrate the advantages of the proposed approach, does not provides insights into the complexity-accuracy trade-off, and does not offers a practical benchmark for classification in explicit content detection.

4 Conclusion

This research study presents the development of a content filtering model designed specifically for streaming education platforms. The model utilizes a combination of computational intelligence techniques to automatically and instantly filter out violent and adult content. A unique aspect of this proposed system lies in its incorporation of deep learning, fuzzy logic and advanced language models, marking the first instance of such integration in the existing literature. This distinctive approach facilitates a comprehensive analysis of digital content by considering both visual and fuzzy-linguistic aspects.

To assess the effectiveness of our model, extensive experiments were conducted employing an artificial dataset. The results demonstrated that our proposed model surpasses a baseline keyword-based filtering method in terms of performance. Our model achieved higher levels of accuracy, precision, recall, and F1-score, substantiating its proficiency in identifying and filtering inappropriate content. Furthermore, the model exhibited rapid content processing, ensuring real-time analysis and filtering without any detrimental impact on platform users. The efficacy of GPT-3 in accurately generating user-readable warning messages to denote explicit content is notable. These messages provide information about the detected explicit content, its potential impact, and suggestions for users to take appropriate actions.

The outcomes of our research study bear significant implications for streaming education platforms. Implementation of our proposed model can contribute to the creation of safer learning environments, particularly for students, by successfully filtering out violent and adult content. This approach reduces instances of false positives and false negatives, thereby augmenting user satisfaction and bolstering trust in the content filtering capabilities of the platform.

The integration of CNNs and IFS enhances accuracy in explicit content detection by leveraging contextual understanding and effectively handling ambiguity and borderline cases. Additionally, GPT-3's generative warnings

play a crucial role in educating users and promoting safety. Moreover, the system's flexibility allows it to adapt to various content types and keep up with emerging trends. On the other hand, the proposed method is complex and requires specialized expertise. It relies on computationally intensive AI techniques, which may limit its use in resource-constrained environments. Gathering a substantial dataset and ensuring its privacy can be time-consuming and ethically challenging. The system's interpretability is reduced, making it difficult to understand its classification decisions. Using pre-trained models introduces external dependencies and potential licensing costs. The system's hypersensitivity to ambiguity can result in false positives or false negatives, making it challenging to achieve the right balance.

Nevertheless, further research and development are necessary to address the intricacies and challenges associated with content filtering, considering the emergence of new types of inappropriate content. Prominent avenues for future research encompass cross-linguistic analysis to evaluate content filtering models across diverse languages and cultures, multimodal content analysis that integrates multiple modalities for comprehensive filtering, examination of biases and fairness in content filtering models, contextual analysis to enhance filtering based on surrounding content and user interactions, the development of specialized models for deepfake detection, real-time anomaly detection for emerging inappropriate content, analysis of evolving content patterns and the creation of adaptive models, amalgamation of AI-driven filtering with human moderation, privacy-preserving filtering methods, education-specific filtering, user empowerment in customizing filtering, exploration of the ethical implications of content moderation, multilingual content filtering, long-form content analysis, and comprehension of the impact of content filtering on mental well-being. Pursuing these research avenues aims to contribute to the establishment of safer and more secure online environments.

References

[1] C. Cayari, "Popular practices for online musicking and performance: Developing creative dispositions for music education and the Internet," J. Pop. Music Educ., vol. 5, no. 3, pp. 295–312, 2021.

[2] M. Al-Dojayli and A. Czekanski, "Integrated Engineering Design Education: Vertical and Lateral Learning," J. Integr. Des. Process Sci., vol. 21, no. 2, pp. 45–59, Jan. 2017, doi: 10.3233/jid-2016-0024.

[3] S. Barua and D. Talukder, "A Blockchain based Decentralized Video Streaming Platform with Content Protection System," in 2020 23rd International Conference on Computer and Information Technology (ICCIT), Sep. 2020, pp. 1–6. doi: 10.1109/ICCIT51783.2020.9392746.

[4] J. Casebeer, N. J. Bryan, and P. Smaragdis, "Meta-AF: Meta-Learning for Adaptive Filters." arXiv, Nov. 21, 2022. doi: 10.48550/arXiv.2204.11942.

[5] D.-C. Chang, C. Chen, and M. Thanavel, "Dynamic reordering bloom filter," in 2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS), Sep. 2017, pp. 288–291. doi: 10.1109/APNOMS.2017.8094131.

[6] M. A. Al-Gunaid, M. V. Shcherbakov, K. S. Zadiran, and A. V. Melikov, "A survey of fuzzy cognitive maps forecasting methods," in 2017 8th International Conference on Information, Intelligence, Systems Applications (IISA), Dec. 2017, pp. 1–6. doi: 10.1109/IISA.2017.8316443.

[7] A. Bastian, S. Tano, T. Oyama, and T. Arnould, "FATE: fuzzy logic automatic transmission expert system," in Proceedings of 1995 IEEE International Conference on Fuzzy Systems., Mar. 1995, pp. 5–6 vol.5. doi: 10.1109/FUZZY.1995.410015.

[8] M. Cai, Y. Shi, J. Kang, J. Liu, and T. Su, "Convolutional maxout neural networks for low-resource speech recognition," in The 9th International Symposium on Chinese Spoken Language Processing, Sep. 2014, pp. 133–137. doi: 10.1109/ISCSLP.2014.6936676.

[9] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[10] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Prog. Artif. Intell., vol. 9, no. 2, pp. 85–112, Jun. 2020, doi: 10.1007/s13748-019-00203-0.

[11] S. Chen, T. Wang, and X. Li, "Research on the improvement of teachers' teaching ability based on machine learning and digital twin technology[J," J. Intell. Fuzzy Syst., vol. 2020, no. 1, pp. 1–12.

[12] A. Hentout, A. Maoudj, and M. Aouache, "A review of the literature on fuzzy-logic approaches for collision-free path planning of manipulator robots[J," Artif. Intell. Rev., vol. 2022, pp. 1–76.

[13] Q. Rao, B. Yu, K. He, and B. Feng, "Regularization and Iterative Initialization of Softmax for Fast Training of Convolutional Neural Networks,"

in 2019 International Joint Conference on Neural Networks (IJCNN), Jul. 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8852459.

[14] Z. Xu, "Research on software credibility algorithm based on deep convolutional sparse coding," MATEC Web Conf., vol. 336, no. 6, 2021.

[15] Y. Kanzawa and S. Miyamoto, "Generalized Fuzzy c-Means Clustering and its Property of Fuzzy Classification Function, JOURNAL OF ADVANCED COMPUTATIONAL IN$^{\mathrm{TEL}}$LIGENCE AND IN$^{\mathrm{TEL}}$LIGENT INFORMATICS," vol. 25, no. 1. pp. 73–82, 2021.

[16] P. Venkata Subba Reddy, "Generalized fuzzy logic for incomplete information," in 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jul. 2013, pp. 1–6. doi: 10.1109/FUZZ-IEEE.2013.6622305.

[17] A. Si and S. Das, "Intuitionistic Multi-fuzzy Convolution Operator and Its Application in Decision Making," in Computational Intelligence, Communications, and Business Analytics, J. K. Mandal, P. Dutta, and S. Mukhopadhyay, Eds., in Communications in Computer and Information Science. Singapore: Springer, 2017, pp. 540–551. doi: 10.1007/978-981-10-6430-2_42.

Biographies

Yong Yu was born in Henan Province, HN, CHN in 1982. He received his Master's degree in software engineering from the Information Engineering University, China, in 2010. From 2010 to 2017, he was a Lecturer. Since 2018 to now, he has been an associate professor with the Computer Engineering Department, Henan Institute of Economics and Trade. He is the author of five books and more than 20 articles. He research interests include big data and artificial intelligence.

Xiaoguo Yin was born in Henan Province, HN, CHN in 1972. He received his Master's degree in economic and strategic science from the University of Zhengzhou, China, in 2008. From 2008 to 2011, he was a Lecturer. From 2012 to 2017 he was an associate professor. Since 2018 to now, he has been a professor with the Management Teaching Department, Henan Institute of Economics and Trade. He is the author of five books and more than 20 articles. He research interests include management and strategic modeling theory.