Mesopotamian Journal of Cybersecurity
Vol. 5, No.2, **pp**.822-841
DOI: https://doi.org/10.58496/MJCS/2025/047; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Research Article

Anti-Cyber Childhood Exploitation: An Online Game Chat Monitoring System

Saja J. Mohammed[1], *, Awos K. Ali[2], Ibrahim M. Ahmed[3]

[1] *Computer Science Department, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq*

[2] *Computer Science Department, College of Education for Pure Science, University of Mosul, Mosul, Iraq*

[3] *Department of Cybersecurity, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq*

ARTICLEINFO

Article History

Received 7 Apr 2025
Revised 19 Jul 2025
Accepted 9 Aug 2025
Published 21 Aug 2025

Keywords

Cybersecurity
Online chatting
Network Systems
Cyber Grooming
Cyberbullying
Optimism and pessimism
LSTM

ABSTRACT

Despite its revolutionary benefits, the Internet has been utilized to abuse children through chat in online gaming. Exhibiting harmful content can negatively impact children's psychology and behaviour, particularly during their developmental years. This paper examines the psychological effects of online predation and the growing risks posed by Internet predators (with focus on children under 15 years old). This paper proposed an Anti-Cyber Childhood Exploitation (A2CE) system, a comprehensive framework designed to detect and prevent three major forms of online abuse: psychological manipulation, cyberbullying, and online grooming. Leveraging advanced Natural language processing (NLP) techniques, A2CE analyses online conversations in real time. The system is trained on three well-known datasets: PUBG, Dota 2, and PAN12. Upon detecting an attack, A2CE provides immediate alerts and warnings to parents, helping mitigate psychological harm. The experimental results demonstrate high detection accuracy: 96.8% for grooming, 94.2% for psychological manipulation, and 92.8% for cyberbullying. These findings indicate that A2CE can be considered a powerful tool for protecting children in this critical age.

1. INTRODUCTION

The Internet is considered to be one of the great accomplishments of the modern digital world, which drastically changed the lives of people regardless of where they lived. The use of the Internet is increasing daily, and this is becoming more popular with the development of network infrastructure. While the Internet has made life more convenient, it can also be a perilous environment, particularly for children. As technology becomes more integrated into our daily lives, so do the threats posed by offenders[1], [2], [3].

Various online messaging systems, such as chat features and even instant messaging applications on social networking websites, have emerged as alternatives to traditional forms of communication. These platforms facilitate peer-to-peer communication without explicit content monitoring, which can be misused for various malicious intents. Examples of harmful activities enabled by the Internet include online extremist groups that normalize dangerous behaviours, spammers that disseminate false information, and those that have significant negative impacts on users' mental health. Targeted chats can be employed to spread hate, plan criminal or terrorist activities, influence victims with propaganda, radicalize individuals, and, in the worst cases, target minors and children for abuse. Unlike public chats or discussions, targeted communications often exploit existing online relationships among group members in the network [4], [5], [6].

Child exploitation is one of the most serious violations of children's rights globally, encompassing various forms of abuse that exploit a child's vulnerability, with the sexual gratification of the abuser or a third party. The consequences for victims can be devastating, both immediately and in the long term, constituting a gross violation of human rights. Child trafficking, sexual exploitation, cyberbullying, and online exploitation are among the many ways children can be victimized.

**Corresponding author. Email: sj_alkado@uomosul.edu.iq*

In recent years, millions of documented cases of online child exploitation have been reported, according to the National Center for Missing & Exploited Children (NCMEC) [5], [7].

A thorough understanding of predator behaviour is essential for developing a reliable and automated surveillance system that enhances children's safety on Internet platforms. Better detection methods are made possible by understanding the patterns used by online predators, which can improve detection methods and educate children on how to respond appropriately in risky situations. However, operational evidence is necessary for digital forensic cases, which require the examination of vast amounts of data and add to the workload of forensic investigations [4], [8], [9].

Recently, various technologies have been proposed to combat child exploitation, albeit with limitations. One of the earliest solutions involved Internet monitoring to scan emails and chat rooms for inappropriate conversations while blocking messages sent to children [10]. Consequently, there is a growing consensus that education is the most critical preventive action that can be taken. Currently, several technologies assist in preventing such circumstances. Major social networks have implemented AI technologies and tools to monitor and identify potential child exploitation. For example, one platform introduced a technology in 2021 that instantly disables any account suspected of distributing harmful content. The United Nations and Europe regularly collaborate with child exploitation helplines to report potentially harmful actions. Additionally, various software solutions and applications are available for users to install on their devices to prevent harmful online activities, such as parental control applications [11], [12]

This paper aims to mitigate online child exploitation, which is implemented to detect and prevent cyber grooming, cyberbullying, and psychological manipulation from occurring during online chatting. The paper is an announcement of the emerging threat of the cyber predation of children by means of online gaming. This highlights the fact that these types of exploitation can pose serious psychological and emotional damage to young people, particularly to people below the age of 15.

One of the main areas of artificial intelligence (AI) that helps address this problem is natural language processing (NLP), which has the ability to teach computers to understand, interpret, and answer human language. NLP can almost translate what people say and communicate with computerized life forms, which facilitates more natural interactions with computers and technologies. NLP is becoming more significant in multiple domains, including customer care, healthcare, finance, and education, as AI technology is still developing some of its strongest applications are text classification, machine translation, information retrieval, sentiment analysis, and the creation of chatbots and virtual assistants[13].

NLP implements various important techniques to analyze or process text. Tokenization examines a piece of text as small print words, sentences or characters, which makes it simpler to analyze. Sentiment analysis allows us to understand the emotional character of a message, whether it is positive, negative or neutral, and how these data can be applied, especially to social media monitoring or product reviews. Named entity detection (NER) is a method of identifying and marking significant structures in a document, such as the name of a person, place, organization, and place as well as the date. The other critical method is Language Modelling, which forecasts the probability of a combination of words, assisting in a variety of tasks, such as speech, text generation, or translation. These techniques are based on neural networks and models and statistical approaches, including N-grams [13], [14].

This paper proposes **the anti-cyber childhood exploitation (A2CE)** system to address these threats, which is an AI-based solution based on NLP checking of online conversations in real-time mode to analyze them and detect various kinds of threats. The aim of A2CE is particularly to detect negative actions such as online grooming, online bullying, and psychological control. When this type of threat is detected, the proposed system immediately alerts children and their parents so that they can resort to efforts to provide protection. Since A2CE is trained on a variety of data, such as PAN12 (overall predator detection), PUBG (emotional tone, specifically chat conversations in gaming), and Dota 2 (cyberbullying), it has the ability to detect many online exploitations and combat them with perfect precision.

2. RELATED WORKS

Chatting over the Internet has attracted the curiosity of many hackers who want to take advantage of children; therefore, many researchers are also developing techniques to prevent cyberattacks. Some articles that propose some measures to counter the online chat threats are as follows:

In [15], an automatic solution to identify grooming discussions in chat messages in the online world based on AI technologies to automate this process was proposed. The training data were drawn via actual conversations via algorithms such as fuzzy-rough feature selection and fuzzy twin support vector machines. Various classifiers were applied to the system to identify the accuracy of the system, and the Gaussian naive Bayes classifier had the highest accuracy of 58.33% with the use of the feature extraction method of bag of words (BoW) and a specific normalization method.

In 2020 [16], an effective method for detecting abusive and bullying online remarks was via machine learning (ML) combined with NLP. Two popular text representation methods, term frequency-inverse document frequency (TF-IDF) and bag-of-words (BoW), are applied to estimate the accuracy of four various machine learning models. The results indicate that the combination of NLP and ML technologies considerably enhanced cyberbullying detection in social media networks. Proper choice of the side features and the functioning of the algorithm had a significant impact on maximizing accuracy.

In a related work conducted in 2021 [17], a convolutional neural network (CNN) was designed to label online conversations on the basis of their content. Since it was performed manually, the Cohen kappa score between the annotators was 0.78742, indicating strong sensitivity. A set of ground truth data (to define the presence of predatory behaviour indicators) was proposed specifically with the CNN model in mind to classify statements that could be described as predatory. When the predator profiles were classified, the model obtained F0.5 = 0.79 and F2 = 0.98. Nevertheless, its effectiveness in a narrower task of identifying certain messages aimed at the process of grooming was lower, with F2 = 0.61. The effects of adding other features, including sentiment analysis, to further enhance the detection accuracy were analysed. The CNN was able to detect predator profiles; however, difficulties were experienced in detecting grooming lines. The article also proposed the use of functionalities of multiple models to improve performance, which is another new method for identifying online predators through modern machine-learning tools.

In 2023 [18], a framework for extracting features via a contrastive learning framework for sentences and effectively managing a conversation with misspellings via sub word data was proposed. An important goal was to obtain a high true positive detection rate to avoid false condemnation by innocent people. The authors implemented a combination of robustly optimized bert approach (RoBERTa) encoders with a supervised Simple Framework for Contrastive Sentence Embedding (SimCSE) model to train support vector machine (SVM) classifier. The experiments indicated promising results, including an F0 score of 0.96, an F1 score of 0.96, and an accuracy of 0.99 in predatory conversation detection, which is a new standard in the domain. Additional trials with different fusion methods revealed that sum fusion of all the configurations produced an accuracy of 0.99, an F1 score of 0.97 and an F0.5 score of 0.98.

In 2024 [19], a new approach was proposed, that is able to strengthen the detection of threats in dynamic environment by employing actor significance thresholds (ASTs) and message significance thresholds (MSTs). In this strategy, the focus was on using high-performance model language, such as bidirectional encoder representations from transformers (BERT) and RoBERTa, to conduct message-level processing. Additionally, it uses a context-sensitive classification method to evaluate actor interactions. The proposed technique benefits from recognizing the multidimensional and dynamic character of current threats, which makes the detection mechanisms more precise and robust. The advantages of this strategy (in terms of resilience and flexibility) were confirmed via cross-dataset experiments, which proved its usefulness in different contexts.

Additionally, in 2024 [20], a study introduced chat-based dynamic difficulty adjustment (ChatDDA), an original structure that assesses the perceived difficulty of players in gaming settings according to the chat communicated in-game. The authors adapted the pretrained language models (RoBERTa, BERT, and Twitter-RoBERTa) jointly with the chat logs of the player un known battlegrounds (PUBG) game to extract the semantic features as a sign of the players being either pessimistic or optimistic with respect to their win. These data were subsequently fed to a feed-forward neural network training, and a high test accuracy of 0.953 was attained. The findings illustrate that ChatDDA was accurate in estimating perceived difficulty in the game, thus improving user engagement, immersion, and experiences inside the game in traditional games and metaverse platforms.

Later, [21],in another article, an effective methodology to improve knowledge identification of cybersecurity via large, unstructured data through chat was proposed. To solve the difficulty of asynchronous and intersecting conversations, the authors designed a heuristic algorithm for BERT-next sentence prediction to break down and disentangle chat threads into coherently constructed dialogues. Such preprocessing greatly increased the accuracy of the following relation extraction model, which was specifically trained to work within the field of cybersecurity. The model was trained and evaluated via a domain specific, dialogue-based relation extraction dataset, advancing the ability to derive actionable intelligence from informal online discussions. The experimental results indicated that the average F1 scores for the dialogue relation extraction task and the thread disentanglement task were 88.4 and 74.9, respectively. Table (I) summarizes the previous works.

TABLE I. SUMMARY OF THE MOST RELATED WORKS

Ref.	Used Method	Contribution	Limitation
[15]	• Feature Extraction: BoW and TF-IDF • Feature Selection: Fuzzy-Rough Feature Selection (FRFS). • Classification: Fuzzy Twin Support Vector Machines (SVM). • Evaluation: 10-Fold Cross-Validation.	• Development of an Intelligent Detection System. • Use of AI Technologies • Diverse Training Data • Feature Extraction and Selection • Addressing Investigator Challenges.	• Dependence on Training Data • Time-Consuming for Data Collection. • Feature Selection Challenges. • Limited Real-World Testing • Evolving Nature of Grooming Language. • Psychological Impact on Investigators.
[16]	• Natural Language Processing (NLP). • Machine Learning Algorithms. • Feature Extraction Techniques: • BoW and TF-IDF.	• Development of Detection Techniques. • Integration of Machine Learning and NLP. • Evaluation of Multiple Algorithms. • Feature Extraction Insights • Addressing a Critical Social Issue.	• Dataset Limitations. • Algorithmic Bias. • Contextual Understanding. • Real-Time Detection Challenges. • Limited Scope of Analysis.
[17]	• Two-Step Classification Approach: • Suspicious Conversation Identification (SCI). • Victim from Predator Disclosure (VFP). • Convolutional Neural Networks (CNNs) • Multilayer Perceptron (MLP). • Line Feature (LiF).	• CNN-based Lexical Feature Extraction • Gold Standard for Line Identification. • Behavioral Feature Analysis.	• Contextual Ambiguity. • Feature Generalizability. • Data Imbalance. • Manual Annotation Challenges. • Platform Specificity.
[18]	• Contrastive Learning Framework • Sentence Embedding with SimCSE • Use of Pretrained Models (RoBERTa). • SVM.	• Effective Use of Sentence Embeddings (a Simple Contrastive Sentence Embedding (SimCSE). • Integration of Pretrained Models (RoBERTa) • Addressing Imbalanced Datasets.	• Imbalanced Dataset Challenges. • Context Sensitivity. • Limited Data Availability. • Potential for Overfitting. • Cultural Differences.
[19]	• Transformer-Based Models (BERT, RoBERTa) • Actor Significance Thresholds (AST) and Message Significance Thresholds.	• Dynamic Thresholding. • Robustness Across Datasets. • Practical Implications.	• Contextual Ambiguity. • Computational Costs. • Ethical and Privacy Concerns. • Generalizability (Focused on Peer-to-Peer Chats).
[20]	• ChatDDA Methodology • Pretrained Language Models: The authors utilize three pretrained language models—BERT, RoBERTa, and Twitter-roBERTa. • Feed-Forward Neural Network	• Novel Methodology called ChatDDA • Custom Dataset Development.	• Dependence on Chat Content. • Potential for Misinterpretation. • Computational Complexity.
[21]	• Heuristic Algorithm Development based on BERT • Thread Disentanglement.	• Innovative Triple Extraction Approach. • Heuristic Algorithm for Dialogue Processing. • Tailored Relation Extraction Model.	• Contextual Understanding. • Scalability Issues. • Focus on Specific Threats.

In contrast to earlier systems that concentrated on individual components, this paper proposes the A2CE system, which distinguishes itself by providing a comprehensive framework that incorporates the detection of psychological manipulation, cyberbullying, and online grooming. Its real-time detection capabilities address a crucial gap in previous methods by enabling prompt notifications and responses. After training on three other datasets—PUBG, Dota 2, and PAN12—A2CE improves performance in various online environments. Moreover, it provides tailor-made security strategies to suit the demands of children below the age of 15. The proposed system immediately alerts parents upon detecting harmful activity, significantly enhancing parental control. Through the use of sophisticated natural language processing methods, A2CE achieves a more nuanced understanding of unsafe conversations. Collectively, these characteristics make A2CE an effective tool for protecting children online.

3. ONLINE CYBER BEHAVIOUR ATTACK PATTERNS

This paper focuses on three types of attacks that can target children under 15 years of age in online game chat, directly affecting their psychological state:

3.1 Cyberbullying

Cyberbullying involves using electronic communications to harass, threaten, or embarrass a child, often through messaging applications, social media sites, or online gaming environments. The anonymity provided by the Internet may prompt bullies to target victims without immediate consequences. Research indicates that victims of cyberbullying may experience severe mental distress, including anxiety, depression, and declining academic performance. Approximately 20% of the students reported having been victims of cyberbullying, highlighting the need for robust preventive measures and safe spaces for affected children. Peers, parents, and teachers play crucial roles in addressing this problem by raising awareness and fostering open discussions about online behaviour [22], [23].

Cyberbullying, grieving, chat spamming, and bug abuse are examples of antisocial or problematic behaviour (often referred to as "toxicity" within the gaming community), including harassment of minorities or specific races [24]. The effects of cyberbullying can be devastating, leading to feelings of harm, humiliation, anger, depression, and even suicidal thoughts. Up to 43% of professionals working remotely have experienced cyberbullying or online harassment, and nearly half of teenagers in the United States have encountered these forms of abuse [25]. Fig. (1) shows the summary of thirteen different studies conducted between 2007 and 2023, according to the cyberbullying research center in 2024 [26].

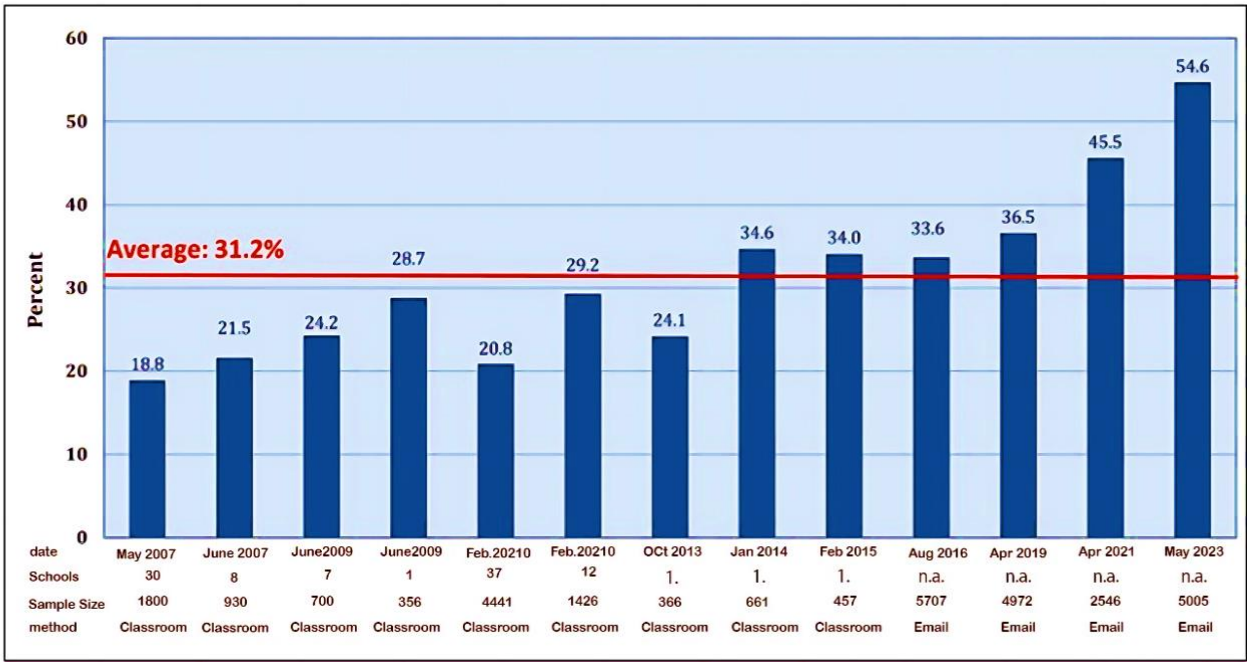


Fig. 1. The lifetime cyberbullying victimization rates [24]

3.2 Online Grooming

The widespread use of social media has significantly facilitated the crime of online child grooming, which was the highest reported crime in the United Kingdom (UK) between 2009 and 2010, according to the Child Exploitation and Online Protection Agency. Online child grooming involves approaching, convincing, and involving a child—the victim—in sexual conduct through the Internet. The perpetrator seeks to establish an emotional and sexual bond with the victim. Sexual grooming can occur in various settings, including online environments, where offenders may create false identities to gain the trust of their children. Consequently, the psychological, physical, emotional, behavioural, and psychosocial aspects of the victim's life may all be affected [27], [28]

To uncover this type of crime, investigators typically analyse conversation texts for grooming tendencies. However, online sexual grooming presents unique challenges, including the fact that communication is recorded and more accessible to third parties. Online predators often rely on emotional manipulation and coercion, targeting children aged 12 to 15 through social networking sites and chat rooms. Common characteristics of grooming include seduction, gradual involvement, and nonviolent coercion [29], [30].

3.3 Optimism and pessimism

Optimism and pessimism are psychological mindsets that influence well-being and shape individuals' perspectives on the world and the future. Pessimism involves expecting negative outcomes, whereas optimism involves anticipating positive outcomes. Individuals' levels of optimism and pessimism can fluctuate over time and across different aspects of life, indicating that these attitudes exist on a continuum [31]

Research shows that children and adolescents who exhibit higher levels of optimism and lower levels of pessimism tend to have better health outcomes, regardless of chronic health issues. Optimism has been linked to a range of positive psychological and physical health outcomes. It contributes to stronger coping skills in the face of adversity, lower incidences of substance abuse, and a decreased risk of cardiovascular and metabolic disorders. Compared with their pessimistic peers, children with an optimistic outlook are associated with better academic and future career performance, reduced anxiety, enhanced emotional resilience, and a lower likelihood of developing depression when compared to their pessimistic peers [32].

Video games, as potential extrospective or introspective shapers from the perspective of children, can be positive and negative at the same time because they strongly depend on the design of the game, the time that is spent on playing, and on the particular emotional or cognitive demands of the child. On the positive side, video gaming can promote cognitive growth, serve as a means of stress relief, and fulfil certain psychological needs. However, excessive or poorly moderated gaming may contribute to emotional difficulties, a decline in prosocial behaviour, and increased levels of pessimism [33].

4. THE EMPLOYED MODELS

The following subsections outline the AI model employed in the proposed system, detailing its components and how it supports the detection of harmful online behaviour.

4.1 Long short-term memory model

The long short-term memory (LSTM) model is a type of recurrent neural network" type that was created to address the vanishing gradient and exploding gradient issues that arise when learning long-term dependencies. It maintains error signals within each cell via a constant error carousel" (CEC). The architecture includes an input gate, an output gate, and a forget gate, which are subsequently added to allow state resetting. These components constitute a vanilla LSTM unit, enabling it to regulate information flow and retain values over various time periods. Although the vanilla LSTM design is the most widely utilized, it is not always the optimal choice [32]. Fig (2) illustrates the structure of the LSTM [33].

Memory blocks, known as cells, constitute a typical LSTM network. The cell state and the hidden state are the two states that are transferred to the subsequent cell. The primary data flow chain that allows nearly unaltered data movement is the cell state, although some linear transformations may occur. Sigmoid gates are employed to add or remove data from the cell state. A gate is comparable to a layer or a sequence of matrix operations with varying individual weights. Since LSTMs utilize gates to control the memorization process, they are designed to circumvent the long-term dependency issue [34].

The initial step in constructing an LSTM network is to identify information that is superfluous and will not be included in the cell at that phase. The sigmoid function, which takes the existing input (X_t) at time t and the previous output of the LSTM unit (h_{t-1}) at time t − 1, determines this process of detecting and eliminating data. The sigmoid function also determines whether a portion of the previous output should be removed. This gate is known as the forget gate" (or f_t); each number in the cell state, C_{t-1}, is represented by a vector with values ranging from 0 to 1 [35], [36]. This is clear in Eq. (1):

$$f_t = \sigma\left(W_f[h_{t-1}, X_t] + b_f\right) \tag{1}$$

Here, W_f and b_f are the weight matrices and bias of the forget gate, respectively, and σ is the sigmoid function. The next step involves updating the cell state and determining and storing information from the new input (X_t). The sigmoid layer and the tanh layer are the two components of this stage. The tanh function assigns weights to the data that have passed through, determining their level of relevance (−1 to 1), after the sigmoid layer initially decides whether to update or disregard the new information (0 or 1). The new cell state is updated by multiplying the two values. This new memory is subsequently appended to the old memory C_{t-1}, yielding C_t [34] Eq. (2) as follows:

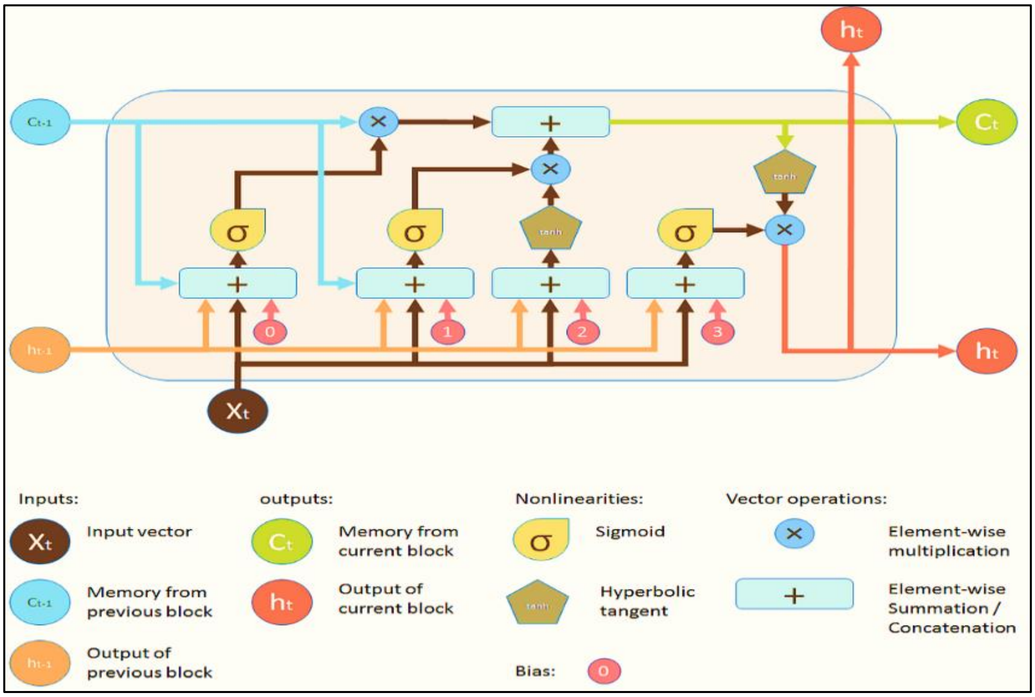


Fig. 2. LSTM structure [35]

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i), \qquad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$
$$N_t = \tanh(W_n[h_{t-1}, X_t] + b_n),$$
$$C_t = C_{t-1}f_t + N_t i_t.$$

The cell states at times t − 1 and t are denoted by C_{t-1} and C_t, respectively, and the weight matrices and bias of the cell state are denoted by W and b. The output values (h_t) in the final step are filtered versions of the output cell state (O_t). The components of the cell state that reach the output are first determined by a sigmoid layer. The new values produced by the tanh layer from the cell state (C_t) are then multiplied by the output of the sigmoid gate (O_t), which ranges from -1 to 1 [35], [36]. See Eq. (3):

$$O_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$h_t = O_t \, tanh(C_t).$$

where W_o and b_o stand for the weight matrices and bias of the output gate, respectively.

4.2 Bidirectional and autoregressive transformers (BART)

Bidirectional and autoregressive transformers (BART) constitute a new transformer framework that includes two best transformer models: a generative pretrained transformer (GPT) decoder and a BERT encoder.

Among the main strengths of the BART (bidirectional and autoregressive transformers) model is its ability to produce texts through containing context in both directions (left-to-right and right-to-left contexts). This two-way processing allows the BART model to generate more fluent, linguistically coherent, and contextually accurate text than other types of transformer models do. BART can be classified as a generative model belonging to the following domains. Bidirectional and autoregressive transformers (BART) are among the latest transformer frameworks that incorporate the two best transformers of the generative pretrained transformer (GPT) decoder and BERT encoder [37], [38].

The bidirectional and autoregressive transformers (BART) model, which is one of its primary strengths, has the ability to generate texts with both left-to-right and right-to-left contexts. BART is a bidirectional processing method that enables BART to produce more fluent, linguistically coherent and contextually fine-tuned text than other types of transformer models of ML and NLP, and it has been successfully applied to a variety of NLP tasks, such as text summarization, machine translation and natural language generation. Owing to its combination of the BERT encoder, which allows the capture of bidirectional context, and the GPT-style decoder, which can generate sequences, the BART can be applied to solve complex multitasks concerning language understanding and generation. While the GPT decoder enables BART to generate text that flows naturally and fits the context, the BERT encoder enhances BART's comprehension of sentence context [37], [38]. Fig.3 shows the base version of the BART model architecture [39].

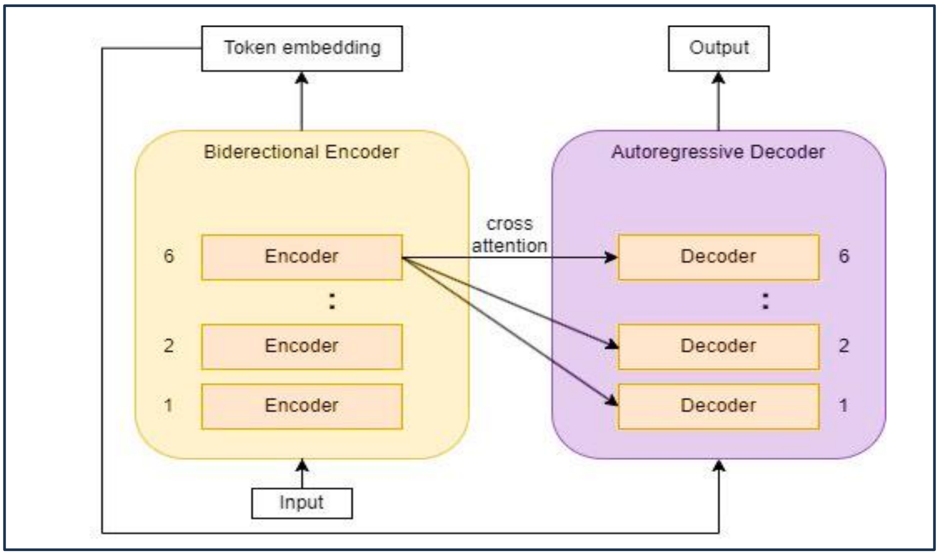


Fig. 3. BART Base version architecture [39].

4.3 Zero-Shot Classification

In the machine learning paradigm known as zero-shot learning (ZSL), a model is trained to identify or categorize items, ideas, or classes that it has never encountered before. ZSL allows models to generalize knowledge from known (seen) classes to generate predictions about unknown (unseen) classes without requiring any labelled instances from those new categories, in contrast to classical supervised learning, which necessitates labelled examples for every class [40]. Flexible and scalable categorization is made possible by zero-shot learning, which is particularly beneficial when working with large datasets that contain many classes. This method also reduces the need for labelled data, which can be costly or time-consuming to gather. Furthermore, ZSL models are more capable of learning and generalizing, potentially improving their performance on subsequent challenges[40].

Zero-shot text classification refers to classifying text into a category without any prior training. This specific type of ZSL application offers the advantage of eliminating the requirement for labels and allows for classifying text into multiple categories via a single model. Additionally, it enables the classification of abstractions without explicitly separating any of the available data for training, which typically requires removing some pertinent and informative data from the testing portion [41].

5. THE PROPOSED ANTI-CHILD EXPLOITATION SYSTEM

The proposed anti-cyber childhood exploitation (A2CE) System (the source code is uploaded to this GitHub repository[1]) focuses on protecting and monitoring children under 15 years old) and mitigating specific types of attacks that can exploit them through online chatting. These attacks can lead to psychological deterioration in children at this critical age. A2CE is trained on three types of datasets, each focusing on a specific type of attack that affects children's mental health through online game chatting. These attacks include cyber grooming, psychological manipulation (optimism and pessimism), and cyberbullying. On the basis of the integrated pretrained model, when a child encounter any of these forms of exploitation through online chatting, the pretrained model detects the attack and executes the prepared actions to defend and protect the victim child directly. Fig. 4 shows the flow diagram of A2CE. The proposed system comprises the following phases.

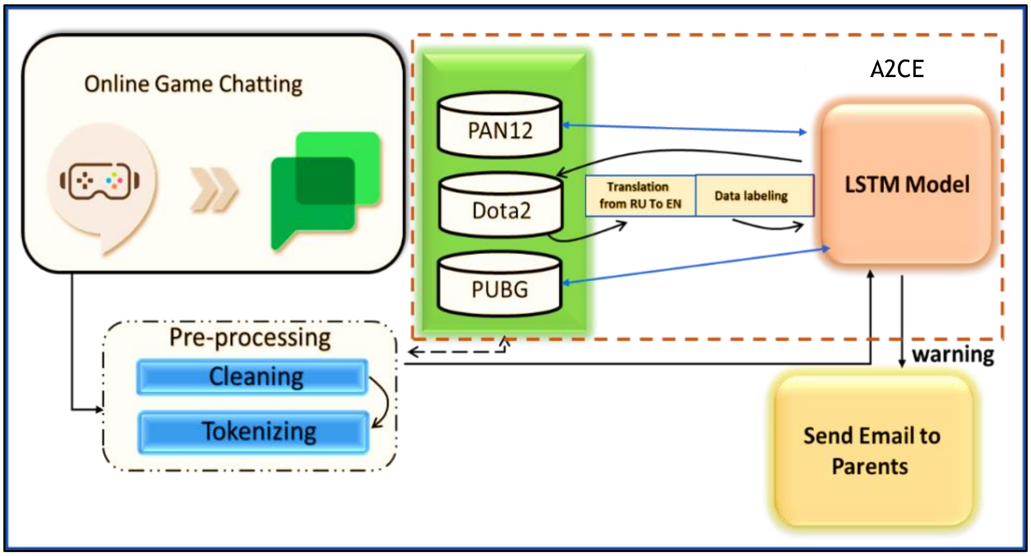


Fig. 4. A2CE flow diagram

5.1 Preparing the datasets:

To achieve the proposed system's goals, three datasets were utilized:

a. PAN12 Sexual Predator Identification (2012)[2]

During the 2012 Conference and labs of the evaluation forum (CLEF), the Publicly Available Natural (PAN) Lab presented a shared challenge in identifying sexual predators. Using information from PJ, the organizers produced a substantial dataset that we refer to as PAN12. They were sampled from IRC channel logs and Omegle[3] site conversation logs. It is particularly challenging to differentiate grooming chats from nonpredatory conversations, as these chats may also involve cybersex between consenting adults. Whenever a conversation was interrupted for longer than twenty-five minutes, the organizers segmented the chats and screened all segments containing more than 150 messages. This process resulted in a total of 222k segments, of which 2.58% were grooming discussions, with the organizers attempting to replicate how grooming is distributed in real online interactions. The segments are divided at a 30:70 ratio between the training and test portions [42].

According to the dataset explanation, the PAN12 task can be separated into two distinct parts:

1. Determine who the predators are among all the users.
2. Identifying which lines or sections of the predator talk best illustrates the predator's inappropriate behaviour [43].

b. PUBG Simulated Chat Messages (Version 1, 2023)[4]

[1] https://github.com/coaka/Grooming
[2] https://pan.webis.de/clef12/pan12-web/sexual-predator-identification.html
[3] Omegle: Talk to strangers! (omegle.com)
[4] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BUE7DW

The "PUBG simulated chat messages" dataset (from the Harvard datasets repository) addresses the challenge of accessing player chat data in the popular game PUBG. It provides researchers with a comprehensive collection of player chat messages categorized into optimism (labelled as 1) and pessimism (labelled as 0) regarding winning and succeeding in the game. The dataset was created through extensive analysis of PUBG's player behaviour, chat system, and current research. Realistic chat conversations were simulated on the basis of gameplay videos and publicly accessible footage. The dataset includes 5,200 labelled messages with linguistic diversity and a representation of various in-game situations. It offers insights into player communication patterns and sentiments.[20]

c. GOSU. AI Dota 2 Game Chats[5](2018)

The Dota 2 dataset includes chat conversations from one of the most well-known eSport disciplines, Dota 2, a video game developed by Valve. This dataset was utilized to train the roflan bot and contains chats from over one million public matchmaking matches. The collection includes chat messages and many characters in multiple languages, with corresponding columns displaying the player slot and the message's timestamp [24], [44].

Preparing these datasets involves several steps, beginning with the collection of appropriate data, followed by cleaning, labelling, validation, and visualization. The PAN12 and PUBG datasets feature binary classification of data, categorized as predator or nonpredator and optimism or pessimism, respectively. Fig. (5) shows the visualization process of the PAN12 and PUBG datasets via the t-distributed stochastic neighbour embedding (T-SNE) tool.

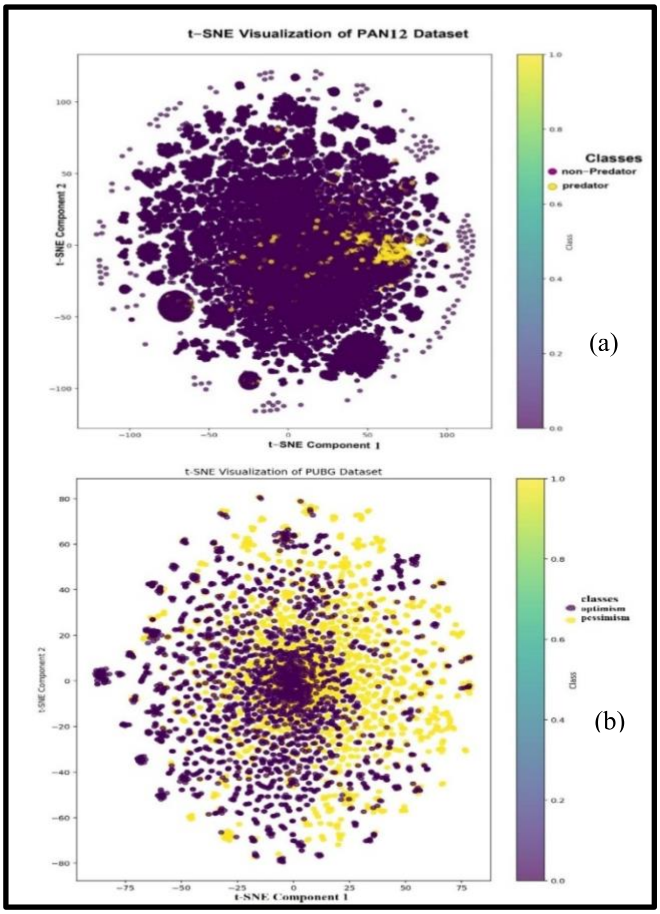


Fig. 5. The T-SNE virtualization step of: (a) PAN12 (b) PUBG

The Dota 2 dataset presents a different case; it is a nonsupervised dataset (not classified), necessitating additional steps for labelling. Furthermore, the Dota 2 dataset is based on the Russian language, as per the Dota 2 game. Consequently, the preprocessing steps began with extracting all conversations from the Dota 2 dataset, translating them from Russian to English

[5] https://www.kaggle.com/datasets/romovpa/gosuai-dota-2-game-chats

(using BART), and subsequently labelling them (using a zero-shot classifier). In general, converting the conversation set to a supervised one involves the following steps, assuming seven different classes in the dataset, as explained in Table (II):

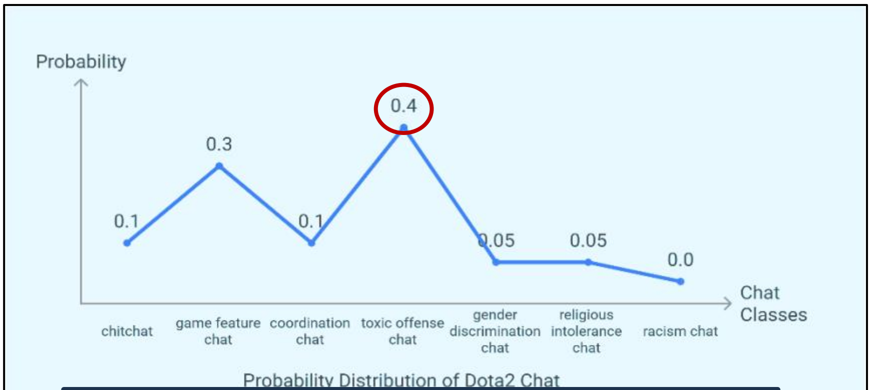
For each record in the selected conversation set:

- On the basis of the meaning of the conversation within the record, the probability of a record belonging to the assumed classifications is determined, ensuring that the sum of all probabilities equals 1.
- The maximum determined probability is selected as the class of the record.

Fig. (6) shows an example of probability values associated with one of these records. The figure indicates that the maximum probability is 0.4; therefore, the record is classified into a “toxic offense chat” class. The above steps were applied to 30,000 records chosen from various sections of the Dota 2 dataset and subsequently to 1,000,000 conversation records. Fig. (7) illustrates the virtualization step of Dota 2 classes after labelling them.

TABLE II. DOTA 2 CLASSES AND THEIR EXPLANATIONS

Chat Type		Description
1.	Chit Chat	Casual discussions that is not serious and informal.
2.	Game Feature Chat	Discussions centered on specific game features, gameplay, and updates.
3.	Coordination Chat	Communication among team members in a game aimed at planning and organizing activities.
4.	Toxic Offense Chat	Damaging and negative communication, such as insults or disparaging comments about others.
5.	Gender Discrimination Chat	Discussions exhibiting gender-based bias or prejudice, often reinforcing stereotypes.
6.	Religious Intolerance Chat	Conversations demonstrating prejudice or hostility toward specific religious communities.
7.	Racism Chat	Discussions involving bias, hostility, or discrimination against individuals based on their race.



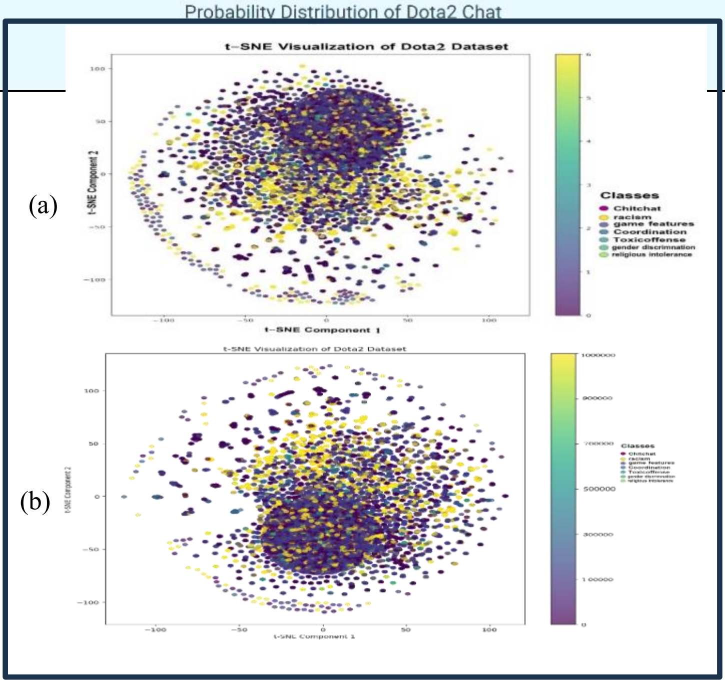


Fig. 7. The T-SNE virtualization results of Dota 2 dataset:
a: for 30,000 records b: for 1,000,000 records

To understand the Dota 2 dataset's preparation and tokenization process in detail, the algorithm first preprocesses the dataset to handle missing values and initializes a translation model from Russian to English. To maximize computational efficiency, text inputs are processed in batch chunks and condensed to 100 characters. Without task-specific fine-tuning, the translated English outputs are then categorized into seven candidate categories via a zero-shot BART-based pipeline. These categories include nontoxic (such as game features and coordination) and toxic (such as racism and gender discrimination) labels. The most likely label is assigned to each chat message, and the results are saved for later analysis. Although input truncation restricts contextual granularity, the method eliminates the need for labelled training data and uses translation to handle low-resource language situations. The pipeline demonstrates practical relevance for real-time moderation in multilingual gaming contexts. Algorithm (1) outlines the details of this processing.

Algorithm (1): Zero-Shot Classification with Translation for Toxicity Detection

```
Input:   dota2 Dataset.csv
Output:  Labelled English dataset

Load the Dota 2 dataset file
Fill missing text values with empty strings
Initialize the translation model from Russian to English
Initialize zero-shot classification pipeline with the BART model
function TranslateText (DataFrame df, Integer chunk size)
    Split data into chunks of size chunk size
    for each chunk do
        Apply the translation model to the first 100 characters of each text
        Append translated text to the new column text English_data
    end for
end function
Defne candidate labels:
          {"chitchat", "game features", "coordination", "toxic
          offense", "gender discrimination", "religious intolerance",
          "racism"}
function ClassifyText (DataFrame df)
    for each translated text in text English_data do
        Predict label using the zero-shot classifier
        Assign the highest scoring label to the column label
    end for
end function
Save labelled DataFrame to Dota2 english all.csv
```

Notably, the proposed model is designed to operate in English, given its popularity and prevalence in many online games. The translation of Russian to English in the Dota 2 dataset was performed via a pretrained BART model. Since the Dota 2 dataset is in Russian, any game environment can be translated into any language, allowing for the selection of the most likely conversational type. This contributes to the standardization of the language used in the A2CE model. Consequently, any other language can be translated into English in a similar manner as Russian can be translated into English, with slight modifications to the employed criteria.

5.2 A2CE Training and Testing

The **A**2CE system employs a bidirectional LSTM architecture with k-fold cross-validation, as shown in Algorithm (2). The approach processes concatenated English-translated datasets through integrated text cleaning, tokenization, and sequence padding pipelines. The model uses dual-stacked bidirectional LSTM layers with embedding inputs, dense connections, and dropout regularization to capture complex contextual relationships. The system demonstrates enhanced generalizability across diverse text corpora, with comprehensive metric reporting, including cross-validation stability analysis. Comparative results indicate significant improvement over baseline models in managing class imbalance and preserving semantic relationships through sequential learning.

```
Algorithm (2) Text Classification Using LSTM with Cross-Validation
Input: Multiple CSV files containing labelled English-translated text
Output: Classification metrics (Accuracy)
Load and concatenate datasets
Clean and preprocess the text data
Define class labels and encode them
Apply TF-IDF vectorization and tokenization
Pad token sequences to a fixed length
Initialize a Bidirectional LSTM model with:
    -Embedding layer
    -Two Bidirectional LSTM layers
    -Dense and Dropout layers
    -Sigmoid activation for output
Compile the model using:
    -Loss: categorical crossentropy
    -Optimizer: Adam
    -Metrics: Accuracy
Split data into training and testing sets
Perform k-fold cross-validation:
for each fold do
    Train the model on the training split
    Validate the model on the validation split
    Record accuracy for this fold
end for
Compute average accuracy across all folds
Print evaluation metrics
```

The PAN12 dataset comprises 357,622 chats; 11,350 of these chats are classified as "predators," whereas 346,272 are categorized as "nonders." Of these, 66,973 chats are used for testing, with 2,014 being "predator chats" and 64911 being "nondredator chats." On the other hand, the PUBG dataset contains 5,200 records for training, which are evenly distributed between optimism and pessimism. Thirty percent of these records are allocated for testing.

Finally, the Dota 2 dataset has 21,659,448 conversation records; 1 million of them are chosen for training and testing at percentages of 80% and 20%, respectively.

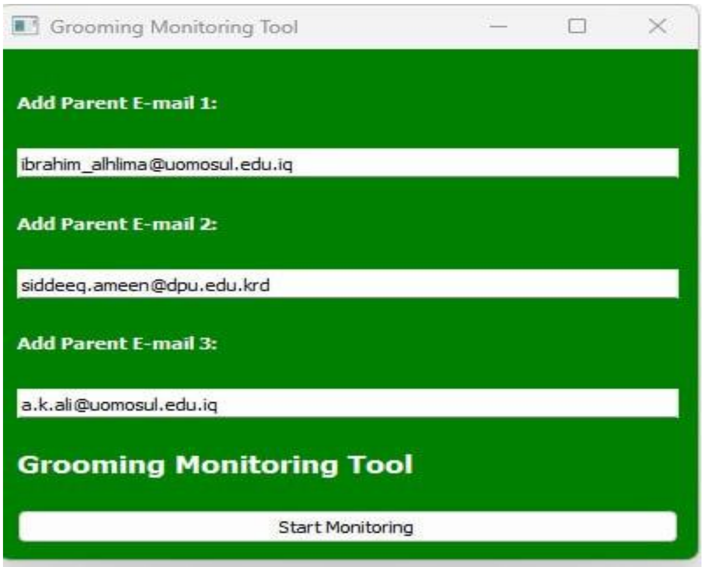
5.3 A2CE application:

A2CE has been completed, encompassing all its aspects and details, via the Python platform, assisted with the "pytorch" and "TensorFlow" packages. Several Python tools are employed, as explained in Table (III).

TABLE III. TOOLS USED

Used Tool	Objective
Tkinter	The majority of Python installations come with this standard Python GUI toolkit.
Bulldozer library	Facilitates running Android applications from a Python environment.
Android Studio	Offers a comprehensive range of features and tools to assist developers build, test, and publishing Android applications.
Game loop	A core component of video game programming that manages the flow of the game, ensuring smooth and consistent operation by repeatedly executing a set of tasks.
NLTK (Natural Language Toolkit)	Provides tools and resources to assist in text analysis, applying machine learning techniques, and handling various language tasks.

The A2CE tool for monitoring grooming behaviours operates within the Android operating system environment by leveraging the game loop simulation platform, which enables Android-based games to run on a computer. The system's code



Grooming Monitoring Tool

Add Parent E-mail 1:

ibrahim_alhlima@uomosul.edu.iq

Add Parent E-mail 2:

siddeeq.ameen@dpu.edu.krd

Add Parent E-mail 3:

a.k.ali@uomosul.edu.iq

Grooming Monitoring Tool

Fig. 8. The designed interface of A2CE application

execution is supported through the use of Bulldozer, a tool that allows Python developers to package their projects for deployment on Android devices. Additionally, the tool uses the NLTK (“Natural Language Toolkit”) library to perform natural language processing on chat conversations within games, enabling the system to interpret linguistic content and more accurately detect inappropriate grooming behaviours.

The proposed A2CE is trained via integrated methods to detect and prevent three types of attacks: sexual online grooming, cyberbullying, and psychological manipulation. A2CE is installed on both the victim's (child’s) and the monitor's (parent's) devices. To increase the accuracy of the monitoring process, A2CE assumes that the

presence of three optional parents to monitor the target child's device. At the start of the monitoring process, the system sends a message to the parent, alerting them to initiate monitoring. This message automatically resents two hours after monitoring begins, serving as a reminder that the child monitoring mode is still active. The application is installed on the target child's device to commence operation, as illustrated in Fig. (8).

When the child is exposed to any of the targeted attacks that the system is pretrained to detect and prevent, a feedback mechanism is activated, altering and sending a message to the parents to warn them that the child is under the influence of a cyberattack that could impact the child's psychological state. To increase security, conversation classification occurs on the target child's device, as this application must function without transferring chat data to a server due to privacy concerns. Fig. (9) illustrates the processing of the A2CE application.

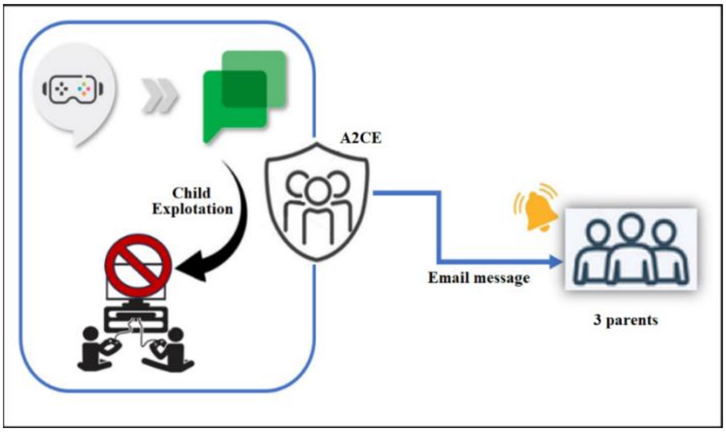


Fig. 9. A2CE application processing

6. PRACTICAL IMPLEMENTATION AND DISCUSSION OF RESULTS

A2CE's practical experiment was conducted on a 13th generation Core i7 with 16 GB of RAM and an RTX (Ray Tracing Texel eXtreme) Nvidia GPU. The practical experiment reveals the confusion matrix results of the proposed system concerning sexual online grooming and psychological manipulation (optimism and pessimism), as shown in Figs. (10) and (11), respectively. Fig. (12) displays the confusion matrix for cyberbullying attacks.

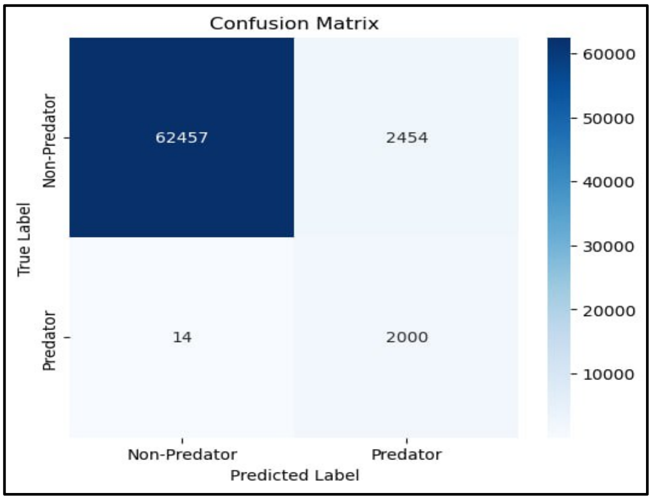


Fig. 10. The confusion matrix result of sexual grooming attack.

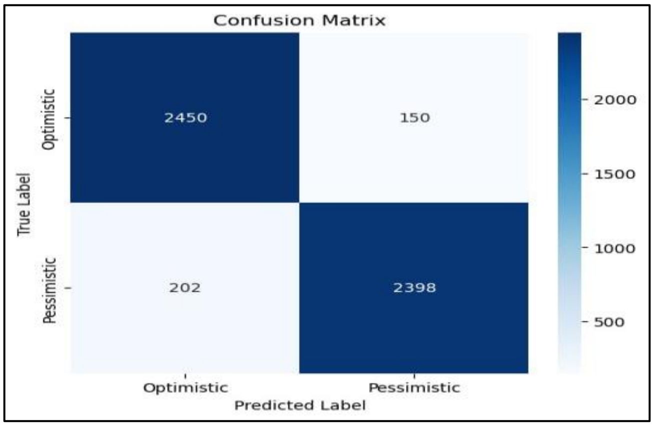


Fig. 11. The confusion matrix result of psychological manipulation attack

The confusion matrix illustrated in Fig. (10) demonstrates strong performance in classifying "nonpredator" instances, with 62,457 true negatives (TNs), indicating that the model correctly identifies the majority of nonthreatening cases. The extremely low false positive (FP) rate of 14 suggests that the model rarely misclassified harmless instances as predators, ensuring high precision for the "nonpredator" class. The presence of 2,000 true positives (TPs) indicates some success in detecting actual predators, although the exact ratio to false negatives (FNs) remains unclear due to the incomplete structure of the matrix. The descending values (60,000--10,000) may represent a threshold analysis, highlighting consistent performance across decision boundaries. Overall, the model excels in minimizing false alarms critical for applications such as security—while maintaining reasonable accuracy for the minority "predator" class. Further clarification of FN values would aid in assessing recall and balance in performance.

Fig. (11) illustrates the classification between the "optimistic" and "pessimistic" categories. The model accurately detects a significant proportion of "optimistic" situations, as evidenced by the high figure of 2,450 true positives (TPs). Similarly, 2,398 true negatives (TNs) demonstrate excellent accuracy in identifying "Pessimistic" cases. The model seldom misclassified "Pessimistic" situations as "Optimistic," maintaining acceptable precision, as indicated by the comparatively low false positive (FP) value of 150. Reliable performance is highlighted by the overall balance between true positives (TPs) and true negatives (TNs), even though the 202 false negatives (FNs) indicate some missed "optimistic" cases. Confidence thresholds may be represented by decreasing numbers (1,000-500), indicating consistent outcomes across various choice limits. According to these measures, the model is particularly adept at reducing overly optimistic predictions, making it suitable for applications requiring careful classification.

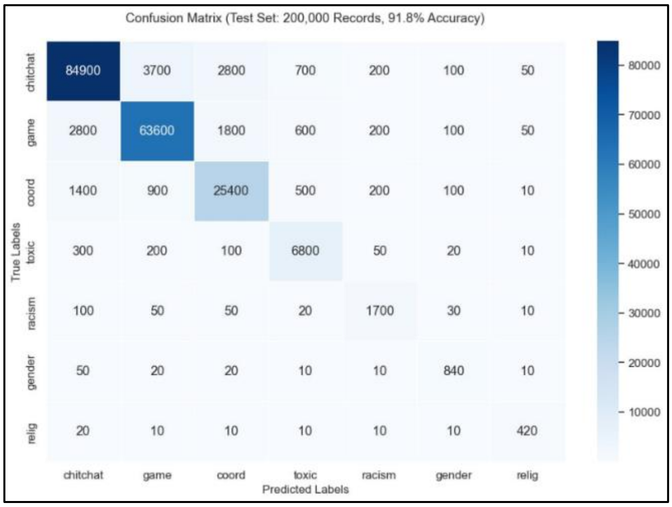


Fig. 12. The confusion matrix result of cyberbullying attack

Finally, Fig. (12) shows outstanding multiclass classification ability across 200,000 test records, with an overall accuracy of 91.8%. Strong true positive rates for each class are displayed by the diagonal numbers (84,900; 63,800; 25,400; etc.), particularly for "critical" (84,900) and "game" (63,800), suggesting that the model is proficient at recognizing these categories. High precision is indicated by the comparatively low off-diagonal values (mainly in the hundreds), which demonstrate minimal misclassification between classes. More common classes (higher values) maintain greater separation than rarer classes do, according to the matrix structure's evident hierarchical pattern. The declining numbers (80,000--10,000) suggest consistent performance across scales and could represent class distributions or confidence thresholds. The model's ability to sustain low confusion rates, even for categories that appear similar, is commendable. The high accuracy, combined with balanced performance across multiple classes, makes the proposed model reliable for complex classification tasks.

The A2CE system meets the standard for deployable AI safety solutions, offering actionable reliability for platforms, schools, or parental controls. Minor refinements could optimize rare class performance, but current metrics already surpass industry norms.

Table (IV) explains the results of the A2CE system using the cross-validation method (k=4), which gives more credibility to AI systems, yielding more accurate and stable results while uncovering issues such as overfitting or underfitting.

TABLE IV. THE PERFORMANCE EVALUATION METRICS OF A2CE (WITH CROSS VALIDATION)

Metrics	**Sexual grooming**	**Psychological Manipulation (Optimism and pessimism)**	**Cyberbullying**
Accuracy	96.8%	94.2%	92.8%
Precision	96.9%	95.3%	95.8%
Recall	99.9%	93.3%	92.8%
F1-score	98.6%	94.3%	94.3%

As shown in Table (IV), the results demonstrate exceptional detection ability in each of the three crucial domains. With a 99.9% recall rate, the system ensures that near-zero missed threats constitute a nonnegotiable requirement for child protection tools. This approach provides a remarkable balance (94.27% F1_score) between identifying threats and preventing false alarms for psychological manipulation. Cyberbullying detection maintains high precision (95.8%), recognizing the majority of genuine occurrences while minimizing false complaints. Across all categories, the accuracy consistently exceeds 92%, demonstrating the system's strong reliability for practical safety applications. These indicators reflect the best-in-class performance in safeguarding individuals against online threats.

The standard deviation was also calculated, as shown in Table (V). A collection of values' standard deviations indicates how dispersed they are in relation to their mean or average. A high standard deviation implies that the values are more widely distributed, whereas a low standard deviation indicates that the data points are closely clustered around the mean.

TABLE V. STANDARD DEVIATION VALUES FOR THE SELECTED METRICS OF THE A2CE MODEL WITH CROSS-VALIDATION

Metric	K fold	Sexual grooming	Standard Deviation	Psychological Manipulation (Optimism, and Pessimism)	Standard Deviation	Cyber-bulling	Standard Deviation value
Accuracy	1	89.477	2.74141682	88.7	1.632812	89.9	0.673846
	2	90.254		90.9		90.89	
	3	93.98		91.78		90.98	
	4	96.2		93.2		91.8	
Precision	1	90.99	3.230715865	87.12	2.775198	89.19	2.034028
	2	95.97		90.25		91.7	
	3	97.2		93.11		92.9	
	4	99.9		94.3		94.8	
Recall	1	90.11	3.79336592	86.89	1.971514	86.14	2.058147
	2	92.14		88.8		89.11	
	3	97.67		90.14		90.12	
	4	99.3		92.3		91.8	
F1-score	1	90.87	3.299457342	87.9	2.050151	89.1	1.515544
	2	96.1		88.7		90.5	
	3	97.9		90.14		91.2	
	4	99.7		93.27		93.3	

As Table (V) explains, the K-fold value increases in all three categories, and all the metrics (accuracy, precision, recall, and F1_score) consistently improve. In terms of category detail, sexual grooming results in the most significant gains, moderate improvements are observed in psychological manipulation, whereas cyberbullying results in the slowest improvements, particularly in terms of recall and the F1 score. Generally, stable performance across various runs or samples is indicated by standard deviation values that are typically low.

The results confirm that the models are most effective at identifying sexual grooming, psychological manipulation, and cyberbullying. The performance metrics consistently improve with increasing K-fold values, indicating that the K-fold value is essential to the model's efficacy. Higher K-fold values should be considered for optimal results, particularly for tasks where the benefits are most pronounced, such as detecting sexual grooming.

A comparative analysis of the results of the proposed model with related work is presented in Table VI.

TABLE VI. THE PERFORMANCE EVALUATION IN OTHER RELATED ARTICLES (IN PERCENTAGES (%))

dataset	Ref	Used Method	Accuracy	Precision	Recall	F1-score
PUBG	[20]	WinOpt-Simple	88.9	88.6	89.8	89.2
		WinOpt-BERT	93.6	92.6	94.7	93.6
		WinOpt-RoBERTa	94.1	94.2	94.4	94.1
		WinOpt-TwRoBERTa	95.3	95.1	95.7	95.3
	The proposed model (A2CE)		94.2	95.3	93.3	94.3
PAN12	[4]	NN (base model)	-	92.3	69.3	79.1
		NN (model-2)	-	93.1	74.9	83.0
		NN (model-3)	-	96.2	74.1	83.7
		Transformer (base model)	-	83.9	65.4	73.2
		Transformer (model-2)	-	90.6	66.6	76.7
		Transformer (model-3)	-	97.8	71.1	90.9
	The proposed model (A2CE)		96.8	96.9	99.9	98.6
Dota 2	[24]	VADER model	82.0	73.0	72.0	71
	The proposed model (A2CE)		92.8	95.8	92.8	94.3

The performance of the A2CE model was evaluated against several existing approaches, as summarized in Table VI. The results reveal that the proposed model achieves impressive accuracy rates of 94.2%, 96.8%, and 92.8% on the PUBG, PAN12, and Dota 2 datasets, respectively. These findings indicate that A2CE consistently outperforms competing models in these benchmarks. Moreover, the model demonstrates superior precision and recall compared with methods such as WinOpt-Simple and WinOpt-BERT, highlighting its ability to effectively balance true positive detection and minimize errors. This improved balance enhances the reliability and robustness of the A2CE model. Additionally, the lower performance metrics observed in rival techniques further underscore the efficacy of the proposed approach. Overall, the A2CE model sets a strong precedent for accuracy and practical applicability in natural language processing tasks.

7. CONCLUSION

The proposed Anti-Cyber Childhood Exploitation (A2CE) system is a major milestone in meeting the protection of children who are exposed to all forms of online abuse, such as grooming, cyberbullying, and psychological manipulation. With the use of natural language processing (NLP) methods and multiple datasets, A2CE achieves high detection accuracies of 96.8%, 94.2%, and 92.8% for grooming, psychological manipulation, and cyberbullying detection, respectively, and exceeds those offered by other previous solutions. The parental alert feature, which is real-time-based, helps eliminate critical flaws in the web-based protection of children that go a long way in preventing harmful exploitation. Additionally, the flexibility of A2CE to a variety of gaming and chat platforms indicates its strength, and the confidentiality of processing in devices enhances secure and ethical deployment. This paper highlights how AI-powered solutions can mitigate digital threat and promote a safer online environment among children through technological advancement and interdisciplinary cooperation. Future projects involve analysing the preparedness to exploit A2CE in the real world, the scale of the model with respect to multi-GPU distributed systems, the extension of the model to use it in a wider field of applications, and the analysis of online voice chat in games to be better able to determine the behaviour of predators.

Conflicts of interest

The authors declare that they have no conflicts of interest to disclose.

Funding

The sponsor and institutional support were not available to the authors.

Acknowledgement

None.

References

[1] Abimbola-akinola and A. Dickson, “The Cyber Crime and Internet and Internet Sexual Exploitation of Children,” *All Student Theses. 107.*, 2017.

[2] K. J. Brakas and M. Alanezi, “Measuring the Extent of Cyberbullying Comments in Facebook Groups for Mosul University Students,” *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 2, pp. 337–348, May 2025, doi: 10.58496/MJCS/2025/021.

[3] K. A. Al-Enezi, S. S. M. Aldabbagh, I. F. T. Al Shaikhli, and J. M. Alwuhaib, “The influence of internet and social media on purchas decisions in Indonesia and a comparison between Indonesia and Kuwait,” *J Comput Theor Nanosci*, vol. 16, no. 3, 2019, doi: 10.1166/jctn.2019.7993.

[4] J. Sekeres, “Methodologies for the Management, Normalization and Identification of Sexual Predation of Minors in Cyber Chat Logs ,” Concordia University, Canada, 2022.

[5] S. H. Abdullah, A. H. Ayad, N. M. Mohammed, and R. M. A. Saad, “Adaptive Fault-Tolerance During Job Scheduling in Cloud Services Based on Swarm Intelligence and Apache Spark,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2, pp. 74–81, Feb. 2023.

[6] M. Al-Dabbagh and A. K. Ali, “Employing light fidelity technology in health monitoring system,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, 2022, doi: 10.11591/ijeecs.v26.i2.pp989-997.

[7] S. Ali, H. A. Haykal, and E. Y. M. Youssef, “Child Sexual Abuse and the Internet—A Systematic Review,” *Human Arenas*, vol. 6, no. 2, 2023, doi: 10.1007/s42087-021-00228-9.

[8] S. A. Baker and A. S. Nori, “Internet of Things Security: A Survey,” in *Advances in Cyber Security. ACeS 2020. Communications in Computer and Information Science*, vol. 1347, Singapore: Springer, 2021, pp. 95–117.

[9] S. J. Mohammed and D. B. Taha, “Paillier cryptosystem enhancement for Homomorphic Encryption technique,” *Multimed Tools Appl*, vol. 83, no. 8, 2024, doi: 10.1007/s11042-023-16301-0.

[10] A. Q. Saeed *et al.*, “Integrating Three Machine Learning Algorithms in Ensemble Learning Model for Improving Content-based Spam Email Recognition,” *Journal of Soft Computing and Data Mining*, vol. 5, no. 2, pp. 188–196, Dec. 2024, doi: 10.30880/jscdm.2024.05.02.014.

[11] S. A. Pasha, S. Ali, and R. Jeljeli, “Artificial Intelligence Implementation to Counteract Cybercrimes Against Children in Pakistan,” *Human Arenas*, 2022, doi: 10.1007/s42087-022-00312-8.

[12] I. M. Ahmed, A. K. Ali, and M. S. Mahmood, “Employing Hybrid Watermarking to Improve Email Security Against Cyber Attacks,” *Journal of Soft Computing and Data Mining*, vol. 6, no. 1, pp. 435–447, Jun. 2025, doi: 10.30880/jscdm.2025.06.01.029.

[13] J. K. Wang, S. K. Wang, E. B. Lee, and R. T. Chang, “Natural Language Processing (NLP) in AI,” in *Digital Eye Care and Teleophthalmology: A Practical Guide to Applications*, 2023. doi: 10.1007/978-3-031-24052-2_17.

[14] K. J. Brakas and M. Alanezi, “A Dynamic DNA Cryptosystem for Secure File Sharing,” *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 2, pp. 424–435, Jun. 2025.

[15] P. Anderson, Z. Zuo, L. Yang, and Y. Qu, “An Intelligent Online Grooming Detection System Using AI Technologies,” in *IEEE International Conference on Fuzzy Systems*, 2019. doi: 10.1109/FUZZ-IEEE.2019.8858973.

[16] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, “Cyberbullying Detection on Social Networks Using Machine Learning Approaches,” in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020*, 2020. doi: 10.1109/CSDE50874.2020.9411601.

[17] S. Preuß *et al.*, “Automatically identifying online grooming chats using CNN-based feature extraction,” in *KONVENS 2021 - Proceedings of the 17th Conference on Natural Language Processing*, 2021.

[18] P. R. Borj, K. Raja, and P. Bours, “Detecting Online Grooming By Simple Contrastive Chat Embeddings,” in *IWSPA 2023 - Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics*, 2023. doi: 10.1145/3579987.3586564.

[19] J. Street, I. Ihianle, F. Olajide, and A. Lotfi, “Enhanced Online Grooming Detection Employing Context Determination and Message-Level Analysis,” *{arXiv preprint arXiv:2409.07958*, Dec. 2024.

[20] M. M. Rezapour, A. Fatemi, and M. A. Nematbakhsh, “A methodology for using players’ chat content for dynamic difficulty adjustment in metaverse multiplayer games,” *Appl Soft Comput*, vol. 156, 2024, doi: 10.1016/j.asoc.2024.111497.

[21] Z. Yang, C. Huang, and J. Liu, “Unveiling Cybersecurity Threats from Online Chat Groups: A Triple Extraction Approach,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2023. doi: 10.1007/978-3-031-40292-0_15.

[22] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, “Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT,” *Information (Switzerland)*, vol. 14, no. 8, 2023, doi: 10.3390/info14080467.

[23] A. Bozyiğit, S. Utku, and E. Nasibov, “Cyberbullying detection: Utilizing social media features,” *Expert Syst Appl*, vol. 179, 2021, doi: 10.1016/j.eswa.2021.115001.

[24] A. K. Singh, “Sentiment Analysis of Dota 2 videogame chat in context of Cyber-bullying,” National College of Ireland, Ireland, 2022.

[25] C. Maurya, T. Muhammad, P. Dhillon, and P. Maurya, “The effects of cyberbullying victimization on depression and suicidal ideation among adolescents and young adults: a three year cohort study from India,” *BMC Psychiatry*, vol. 22, no. 1, 2022, doi: 10.1186/s12888-022-04238-x.

[26] J. W. Patchin and S. Hinduja, “Cyberbullying Research Center,” Feb. 16, 2024. Accessed: Feb. 02, 2025. [Online]. Available: https://cyberbullying.org/

[27] P. R. Borj, K. Raja, and P. Bours, “Online grooming detection: A comprehensive survey of child exploitation in chat logs,” *Knowl Based Syst*, vol. 259, 2023, doi: 10.1016/j.knosys.2022.110039.

[28] T. R. Ringenberg, K. C. Seigfried-Spellar, J. M. Rayz, and M. K. Rogers, “A scoping review of child grooming strategies: pre- and post-internet,” *Child Abuse Negl*, vol. 123, 2022, doi: 10.1016/j.chiabu.2021.105392.

[29] N. Lorenzo-Dus, C. Evans, and R. Mullineux-Morgan, *Online Child Sexual Grooming Discourse*. 2023. doi: 10.1017/9781009314626.

[30] G. M. Winters and E. L. Jeglic, *Sexual Grooming: Integrating Research, Practice, Prevention, and Policy*. 2022. doi: 10.1007/978-3-031-07222-2.

[31] R. Sun, “More than optimism and pessimism: investor emotions and stock returns,” University of Glasgow, 2025.

[32] M. Q. Abdullah, “Optimism/Pessimism and Its Relationship with Locus of Control Among Children and Adolescents,” *Mathews Journal of Psychiatry and Mental Health*, vol. 3, no. 1, 2018.

[33] A. Lobel, R. C. M. E. Engels, L. L. Stone, W. J. Burk, and I. Granic, “Video Gaming and Children’s Psychosocial Wellbeing: A Longitudinal Study,” *J Youth Adolesc*, vol. 46, no. 4, 2017, doi: 10.1007/s10964-017-0646-z.

[34] G. Van Houdt, C. Mosquera, and G. Nápoles, “A review on the long short-term memory model,” *Artif Intell Rev*, vol. 53, no. 8, 2020, doi: 10.1007/s10462-020-09838-1.

[35] X. H. Le, H. V. Ho, G. Lee, and S. Jung, “Application of Long Short-Term Memory (LSTM) neural network for flood forecasting,” *Water (Switzerland)*, vol. 11, no. 7, 2019, doi: 10.3390/w11071387.

[36] J. P. Tan, A. L. A. Ramos, M. V. Abante, R. L. Tadeo, and R. R. Lansigan, “A Performance Review of Recurrent Neural Networks Long Short-Term Memory (LSTM),” in *2022 3rd International Conference for Emerging Technology, INCET 2022*, 2022. doi: 10.1109/INCET54531.2022.9824567.

[37] M. Lewis *et al.*, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.703.

[38] G. Hartawan, D. Sa’adillah Maylawati, and W. Uriawan, “JIP (Jurnal Informatika Polinema) Halaman| 535 BIDIRECTIONAL AND AUTO-REGRESSIVE TRANSFORMER (BART) FOR INDONESIAN ABSTRACTIVE TEXT SUMMARIZATION”.

[39] S. Wah Tan, C. P. Lee, K. M. Lim, C. Tee, and A. Alqahtani, “QARR-FSQA: Question-Answer Replacement and Removal Pretraining Framework for Few-Shot Question”, doi: 10.1109/ACCESS.2023.1120000.

[40] H. Li, F. Wang, J. Liu, J. Huang, T. Zhang, and S. Yang, “Micro-Knowledge Embedding for Zero-shot Classification,” *Computers and Electrical Engineering*, vol. 101, 2022, doi: 10.1016/j.compeleceng.2022.108068.

[41] C. F. Moreno-Garcia, C. Jayne, E. Elyan, and M. Aceves-Martins, “A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews,” *Decision Analytics Journal*, vol. 6, 2023, doi: 10.1016/j.dajour.2023.100162.

[42] M. Vogt, U. Leser, and A. Akbik, “Early detection of sexual predators in chats,” in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2021. doi: 10.18653/v1/2021.acl-long.386.

[43] Y. Li, “Identifying Online Sexual Predators Using Support Vector Identifying Online Sexual Predators Using Support Vector Machine Machine,” *School of Computer Science*, 2020, doi: 10.21427/20ba-8g14.

[44] A. K. Singh, “Sentiment Analysis of Dota 2 videogame chat in context of Cyber-bullying MSc Research Project Masters of Science in Data Analytics.” [Online]. Available: https://www.vuelio.com/uk/resources/white-papers/pr-media-travel-trends-2021/