information

Article

Enhancing Child Safety in Online Gaming: The Development and Application of Protectbot, an AI-Powered Chatbot Framework

Anum Faraz [1], Fardin Ahsan [1], Jinane Mounsef [1,*], Ioannis Karamitsos [2] and Andreas Kanavos [3]

[1] Electrical Engineering Department, Rochester Institute of Technology, Dubai 341055, United Arab Emirates; af1653@g.rit.edu (A.F.); fxa8225@g.rit.edu (F.A.)
[2] Graduate Programs and Research Department, Rochester Institute of Technology, Dubai 341055, United Arab Emirates; ixkcad1@rit.edu
[3] Department of Informatics, Ionian University, 49100 Corfu, Greece; akanavos@ionio.gr
* Correspondence: jmbcad@rit.edu

Abstract: This study introduces Protectbot, an innovative chatbot framework designed to improve safety in children's online gaming environments. At its core, Protectbot incorporates DialoGPT, a conversational Artificial Intelligence (AI) model rooted in Generative Pre-trained Transformer 2 (GPT-2) technology, engineered to simulate human-like interactions within gaming chat rooms. The framework is distinguished by a robust text classification strategy, rigorously trained on the Publicly Available Natural 2012 (PAN12) dataset, aimed at identifying and mitigating potential sexual predatory behaviors through chat conversation analysis. By utilizing fastText for word embeddings to vectorize sentences, we have refined a support vector machine (SVM) classifier, achieving remarkable performance metrics, with recall, accuracy, and F-scores approaching 0.99. These metrics not only demonstrate the classifier's effectiveness, but also signify a significant advancement beyond existing methodologies in this field. The efficacy of our framework is additionally validated on a custom dataset, composed of 71 predatory chat logs from the Perverted Justice website, further establishing the reliability and robustness of our classifier. Protectbot represents a crucial innovation in enhancing child safety within online gaming communities, providing a proactive, AI-enhanced solution to detect and address predatory threats promptly. Our findings highlight the immense potential of AI-driven interventions to create safer digital spaces for young users.

Keywords: artificial intelligence; chatbot technology; child online safety; generative pre-trained transformer (GPT); machine learning; online gaming; predatory behavior detection

Citation: Faraz, A.; Ahsan, F.; Mounsef, J.; Karamitsos, I.; Kanavos, A. Enhancing Child Safety in Online Gaming: The Development and Application of Protectbot, an AI-Powered Chatbot Framework. *Information* **2024**, *15*, 233. https://doi.org/10.3390/info15040233

Academic Editor: Xiao-Fang Liu

Received: 18 March 2024
Revised: 11 April 2024
Accepted: 13 April 2024
Published: 19 April 2024

1. Introduction

The digital landscape of online gaming has evolved dramatically, transitioning from a mere form of entertainment to a crucial aspect of social interaction, especially among children. In the United States, a staggering 90% of children consider online gaming as their favorite pastime [1]. However, the increasing popularity of online gaming has also highlighted significant risks within these virtual environments, particularly for younger audiences [2]. Representing a third of the global internet user base, which totals 4.95 billion, children often engage in online activities unsupervised, exposing them to cyberbullying, harassment, and, more alarmingly, grooming and pedophilia [3,4].

Research has emphasized the severity of these threats, with a notable proportion of child abuse cases being traced back to interactions on social media, video chat rooms, and gaming platforms [5]. The pressing need to protect vulnerable children online has driven the search for effective early detection and intervention strategies against potential predators.

The discourse on child safety online is increasingly focusing on comprehensive solutions that address the complex nature of digital interactions [6]. Traditional protection

methods, often simplistic, fail to match the pace of the evolving and intricate online gaming ecosystem. The advent of AI and machine learning (ML) technologies offers promising new avenues for enhancing online child safety. However, these advancements come with their own set of challenges, including ethical concerns and the need for strict regulatory oversight [7].

Significant strides have been made in the application of AI and ML for detecting predatory behavior in digital communications [8–29], despite challenges such as data scarcity and the lack of real-time application in gaming contexts [4].

In this study, we present Protectbot, a pioneering chatbot framework specifically designed to enhance the safety of children within online gaming platforms. Protectbot utilizes DialoGPT, an advanced transformer-based conversational AI model, to simulate child-like interactions for the early detection of potential predatory actions through real-time chat analysis. This innovative approach enables Protectbot to integrate seamlessly into gaming environments, significantly surpassing previous models in predator detection accuracy by leveraging a novel text classification strategy trained on the PAN12 dataset [30].

The contributions of our research are manifold:

- The development of Protectbot as a robust AI-driven framework for the real-time detection of predatory behavior, aimed at safeguarding children in online gaming environments.
- The achievement of superior detection performance through extensive training and evaluation on the PAN12 dataset, marking a significant advancement in AI-based child protection methodologies.
- The validation of our model's effectiveness and its practical applicability, as demonstrated by its performance on a unique dataset of 71 predatory conversations sourced from the PJ website [22].

In addressing the need to protect children from online predators, especially within gaming platforms, current methodologies primarily analyze chat logs post-interaction, a process that inherently delays intervention and poses potential risks to child safety online [2]. This recognition highlights the urgency for a shift towards proactive detection mechanisms that can identify early indicators of predatory behavior, offering more immediate and effective protection for minors engaged in online gaming. The innovative Protectbot system sets a new standard in the fight against online child predation, marking a significant advancement in child protection efforts.

In the quest to shield children from online predators, particularly on gaming platforms, current methodologies predominantly focus on analyzing chat logs for predatory behavior [2]. This approach, while valuable, operates within a limited scope, primarily conducting post-chat analysis. Such a strategy inherently delays intervention, posing potential risks to children's safety online. Recognizing this limitation underscores the necessity for a shift from reactive measures to proactive detection mechanisms. These mechanisms should ideally identify early indicators of predatory behavior, thereby offering more immediate and effective protection for minors engaged in online gaming. The innovative Protectbot system introduces a forward-thinking solution to the challenge of online child predator detection, setting a new standard in child protection efforts:

- Interactive Chatbot Initiation: Protectbot utilizes an interactive chatbot as the initial interface with users. This novel application of chatbot technology as a preventive tool diverges from the passive, analytical roles chatbots typically play in existing solutions.
- Advanced Conversational Modeling: By integrating the DialoGPT framework, Protectbot achieves a high level of conversational realism. This framework allows for the generation of contextually relevant responses, ensuring the chatbot can engage in meaningful and natural dialogues. This capability is critical for maintaining engagement without arousing suspicion, a distinct advantage over more rudimentary interaction models.

- Sophisticated ML Classification: Following interactions, chat logs are subjected to detailed scrutiny using an advanced ML classification model. This model is finely tuned to pick up on nuanced linguistic and behavioral cues that may suggest predatory intent, demonstrating a significant improvement in identifying potential threats.
- Automated Reporting for Rapid Intervention: Unique among its peers, Protectbot includes an automated reporting feature designed to alert law enforcement agencies swiftly. This functionality facilitates a faster response to identified threats, effectively closing the gap between detection and action, an area where prior research has shown limitations.

This paper is organized as follows: Section 2 offers a comprehensive review of the existing literature on child protection within online gaming environments, highlighting key advancements and identifying gaps that our work seeks to address. Section 3 details the design and implementation of Protectbot, elucidating the system architecture and the underlying ML models that facilitate the detection of predatory behavior. Section 4 presents our experimental results, benchmarking Protectbot's performance against leading methods and discussing the implications of these findings within the context of online child safety. Finally, Section 5 concludes the study by reflecting on Protectbot's pivotal role in enhancing child safety in online gaming spaces and outlines future research directions to further advance the field.

2. Related Work

The urgency to protect children, the most vulnerable demographic in online gaming chat rooms, has led to diverse methodological approaches across academia and industry. This review focuses on three pivotal areas: AI tools for child protection, conversational models for safety, and text classification for predatory behavior detection. These areas collectively embody the multifaceted strategy required to ensure online safety, setting the stage for a nuanced evaluation of their effectiveness and future potential.

The COVID-19 pandemic and subsequent transition to online social activities have notably increased cyberbullying risks among children. Despite significant advancements in AI-driven content moderation on social media and gaming platforms, the lack of transparency concerning these AI tools' development, deployment, and the datasets used for training marks a critical need for more accessible AI solutions to effectively combat cyberbullying [31].

Reflecting global responses to online safety, India's June 2020 ban on 59 Chinese apps, including the widely popular PUBG, underlines a legislative approach to protecting minors' mental health. PUBG's subsequent introduction of AI-driven playtime restrictions and content filtering exemplifies the industry's proactive shift toward ensuring child safety during the pandemic [32].

Amid these technological and regulatory efforts, the "tech4good" initiative stands out, emphasizing the ethical imperatives driving AI development for Child Online Protection (COP) from 2017 to 2022 [33]. This movement highlights the critical role of AI and chatbots in not only safeguarding, but also educating and empowering children in increasingly digitalized lifestyles.

Innovative applications of chatbots like Negobot in detecting online predators showcase the potential of AI to improve child safety [19,24,26]. Employing a dataset from the PJ website, Negobot uses game theory and sophisticated chat analysis for predatory behavior identification. Despite its innovative approach, Negobot's limited testing underscores the challenges in broader applicability and effectiveness, pointing toward the necessity for ongoing improvement and adaptation in AI technologies for child protection.

Moreover, advancements in text classification algorithms demonstrate potential in identifying inappropriate behaviors within chat logs, though challenges remain in interpreting slang and evolving language patterns [34,35]. BotHook's theoretical model lacked empirical testing, underscoring the gap between concept and practical application [36]. Conversely, a chatbot deployed on Omegle, utilizing SVM and multinomial Naive Bayes

classifiers for emotional and opinion analysis, represents a more concrete effort. This initiative, categorizing user intentions and improving through extended studies, indicates progress in real-world applications.

In summary, the field's current state illustrates a dynamic exploration of AI and ML tools to protect children in online environments. Despite the challenges, including data limitations and the need for greater transparency and empirical validation, these efforts underscore a committed trajectory towards leveraging technology for child safety in digital spaces. Future work must focus on refining these technologies, ensuring they are adaptable, ethically grounded, and effectively integrated into online platforms to create a safer digital ecosystem for children.

The exploration into AI's role in combating online child exploitation extends beyond chatbots, with significant innovations like the "Sweetie 2.0" avatar developed by Terre des Hommes in 2013, targeting webcam-based child predators [37]. This AI-driven approach demonstrates the potential of AI in investigative efforts to identify and prevent online child exploitation. However, integrating such AI technologies into the criminal justice system raises legal and ethical challenges, particularly concerning privacy rights and the risk of misuse. These issues emphasize the need for a careful balance between leveraging technological advancements for child protection and safeguarding individual rights and freedoms.

Research on detecting predatory behavior online has largely concentrated on analyzing social media chat logs [38–42]. Despite the significant body of work in this area, there remains a notable research gap in examining such behaviors within online gaming platforms [2]. This gap is largely due to the difficulties in accessing public datasets of online gaming chats, where privacy concerns limit data collection and preservation. Nonetheless, the approach to detecting predatory behavior in chat logs has been framed as a text classification challenge, utilizing various ML algorithms, from support vector machines (SVMs) to convolutional neural networks (CNNs), to classify text data effectively [8–18,20,21,23,25,27–30].

Key datasets for training and evaluating these classifiers include the PAN12 and PJ datasets. Early efforts in this domain utilized basic text classification techniques to distinguish between benign and predatory conversations, highlighting AI's potential in this field [43]. Subsequent research has incorporated natural language processing (NLP) techniques to analyze the linguistic features of chat room conversations more deeply, demonstrating the effectiveness of NLP in identifying specific predatory linguistic cues [44]. The progression of this research area has seen the adoption of deep learning models, such as recurrent neural networks (RNNs), which offer enhanced capability in capturing the contextual and temporal dynamics of textual communication [45].

Significant strides have been made in the field, notably in [28], who harnessed the PAN13 dataset to forge classifiers adept at pinpointing predatory behavior. Their findings revealed that logistic regression models excelled in accuracy, marking a pivotal advancement in the domain [21,28]. This research path has since broadened to encompass an array of classifiers and analytical techniques, underscoring the rapidly advancing capabilities and potential of AI in combating online predatory behavior [9,11,23,25,29,46].

A novel methodology that integrates fastText embeddings with ML classifiers, notably focusing on the SVM classifier, to differentiate between predatory and non-predatory chat logs, was introduced. This approach capitalizes on the established efficacy of fastText across diverse text classification scenarios, utilizing its robust word embeddings that adeptly encapsulate morphological information [47,48]. This aspect is particularly crucial for analyzing informal texts characterized by non-standard language usage.

Further innovation was seen in [10], where the text classification framework was augmented by incorporating textual, behavioral, and demographic features, utilizing both SVM and Bernoulli NB classifiers. A pioneering method that merges digital forensic investigation with ML techniques, including logistic regression, XGBoost, multi-layer perceptron (MLP), and long short-term memory (LSTM), for chat log classification was

introduced in [20]. Kirupalini et al. adopted SVM to identify child grooming behavior, integrating an age detection mechanism via a deep neural network (DNN) to specifically target chats involving children, thus aiming to reduce false positives [18].

The research scope was broadened by Wani et al., who extracted vocabulary-based and emotion-based features, applying them to classifiers such as decision trees (DT), SVM, and random forest (RF) to distinguish between predatory and non-predatory conversations [27]. Preuss et al. implemented a two-step classification process, initially extracting features using CNNs and, subsequently, classifying them through MLP, blending sentiment analysis with lexical features to attain a deeper insight into conversational content [49]. Agarwal et al. unveiled a technique for detecting predatory behavior by leveraging bidirectional encoder representations from transformers (BERT) in conjunction with feed-forward neural networks, demonstrating the sophisticated ability of language models to decode the intricacies of chat room interactions [8].

Collectively, these studies highlight the advancing skill and potential of AI technologies in identifying and countering predatory behavior online, emphasizing the critical need for ongoing research and innovation in this vital domain of cyber safety.

Our investigation contributes to this field by employing fastText embeddings alongside SVM classifiers for the classification of chat logs as predatory or non-predatory. This strategy, previously applied successfully in varied contexts such as topic classification in Indonesian-language tweets, sentiment analysis, and Tibetan text classification, shows considerable promise in addressing the complexities of language and writing errors [47,48,50]. The proficiency of fastText with SVM in navigating language nuances, especially within informal communication channels like chat logs, is attributed to fastText's generation of robust word embeddings by considering each word as a series of character n-grams. This capability is instrumental in capturing morphological details, making it exceptionally suited for analyzing informal texts where non-standard language use is prevalent.

By integrating fastText embeddings with SVM classifiers, this research explores an under-investigated domain, tackling the challenges presented by informal language, abbreviations, and rapid topic shifts prevalent in online communications. This methodological choice not only leverages the strengths of fastText and SVM technologies but also seeks to bridge a significant gap in the current literature on cybersecurity and computational linguistics. The initiative to adapt this advanced approach for detecting predatory behavior in chat logs underscores the potential of AI in enhancing online safety, particularly for vulnerable groups such as children, marking a significant contribution to the discipline.

In the rapidly advancing field of AI, the selection of an optimal model for NLP tasks, such as conversational AI applications, necessitates a nuanced understanding of each model's capabilities and limitations. Recent evaluations, such as [51], have contributed valuable insights into the performance of state-of-the-art models like GPT-4. Their findings suggest that while GPT-4 exhibits unparalleled capabilities in generating human-like text, its efficacy in specific NLP tasks, including text classification, may not consistently surpass that of its predecessors, GPT-2 and GPT-3. This nuanced performance underscores the importance of selecting AI models that align closely with the specific requirements of the application at hand. For tasks that demand not only linguistic realism but also specialized knowledge or nuanced understanding, earlier models such as GPT-2 or GPT-3 may offer a more balanced solution, combining adequate conversational abilities with task-specific effectiveness.

3. Methodology

In this section, we delve into the Protectbot system architecture, a sophisticated framework designed for detecting and reporting potential predatory behavior in online gaming environments. Illustrated in Figure 1, the architecture is delineated through a six-step operational workflow, ensuring comprehensive coverage from initiation to report generation. This workflow diagram illustrates the step-by-step process Protectbot follows to

analyze conversations and generate reports for law enforcement, highlighting the system's comprehensive approach to online safety.

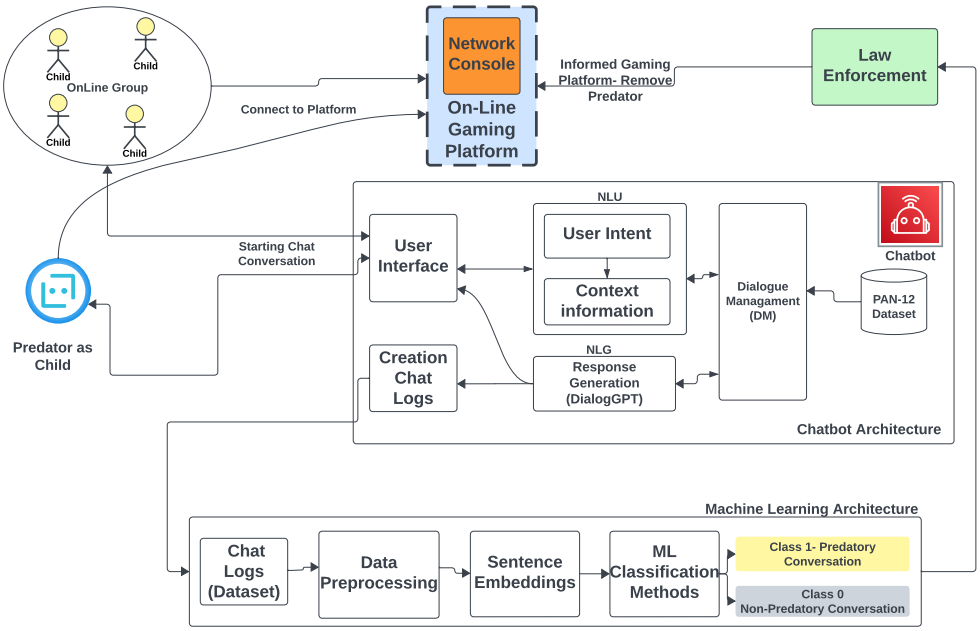

Figure 1. The operational workflow of Protectbot system architecture (6 steps).

- Step 1: Initiation of Chat. The process begins when a user engages in conversation with Protectbot, marking the initial phase of the detection process.
- Step 2: Chat Sent to the Conversational Model. Chats are immediately processed by a conversational model utilizing the DialoGPT framework, selected for its superior ability to generate context-aware, human-like responses. The choice of DialoGPT, trained on diverse internet-based dialogues, ensures the generation of responses that are indistinguishably similar to human interactions, essential for maintaining the predator's engagement.
- Step 3: Generated Response Sent to the Chatbot. Responses crafted by the conversational model are relayed back to Protectbot, which then continues the dialogue with the user. This step is critical for providing a seamless and realistic conversation flow, encouraging the predator to reveal their intentions without suspicion.
- Step 4: Response Sent to the Potential Predator. The conversation progresses with Protectbot sending the generated responses back to the user, meticulously maintaining the dialogue's natural and fluid exchange.
- Step 5: Chat Logs Sent to the Classification Model. Upon concluding the conversation, chat logs are analyzed by an ML classification model. This model scrutinizes the dialogue for specific linguistic and behavioral indicators of predatory intent, such as grooming techniques or attempts to elicit personal information. The model's training emphasizes the identification of nuanced predatory patterns, balancing sensitivity to predatory cues with the necessity to minimize false positives, thereby ensuring ethical considerations regarding privacy and accuracy are upheld.
- Step 6: Report Generation for Law Enforcement Agencies. In instances where predatory behavior is detected, an automated report is generated and dispatched to law enforcement agencies. This report contains detailed insights from the conversation, structured to expedite the authorities' review and action. The collaboration with law enforcement is built on established protocols to ensure rapid and appropriate responses to the identified threats.

This architecture leverages cutting-edge AI and ML technologies to foster a protective environment for children in online gaming platforms. By simulating realistic interactions and analyzing these exchanges for predatory behavior, Protectbot signifies a notable advancement in online child safety measures. The system's integration within gaming environments poses unique challenges, including real-time processing requirements and the adaptability to various gaming contexts. Nonetheless, Protectbot's innovative approach, emphasizing ethical AI use, privacy protection, and efficient law enforcement collaboration, sets a new benchmark in safeguarding children from online predators.

3.1. Conversational Language Generation Model

The core of Protectbot's interaction capability with users is its conversational language generation model, leveraging DialoGPT, a variant of the Generative Pre-trained Transformer (GPT) specifically designed for generating human-like conversational responses. This model is instrumental in enabling Protectbot to simulate engaging and natural dialogues, a critical feature for maintaining the illusion that users are interacting with a child rather than an AI system.

DialoGPT extends the GPT-2 architecture, itself an advancement in language understanding and generation, by focusing specifically on conversational contexts. This specialization allows for more nuanced and appropriate responses in chats, an essential aspect of interacting believably with potential child predators. The architecture operates on a server, processing incoming chat messages via web requests (HTTP REST API calls) to generate responses, which are then posted back into the chat room, facilitating a seamless conversation flow.

Figure 2 illustrates the Protectbot's conversational language generation model architecture. It outlines the steps from receiving a user's message to posting an AI-generated response back into the chat. This process is critical for maintaining real-time interaction, which is essential for the system's credibility and effectiveness in engaging potential predators.

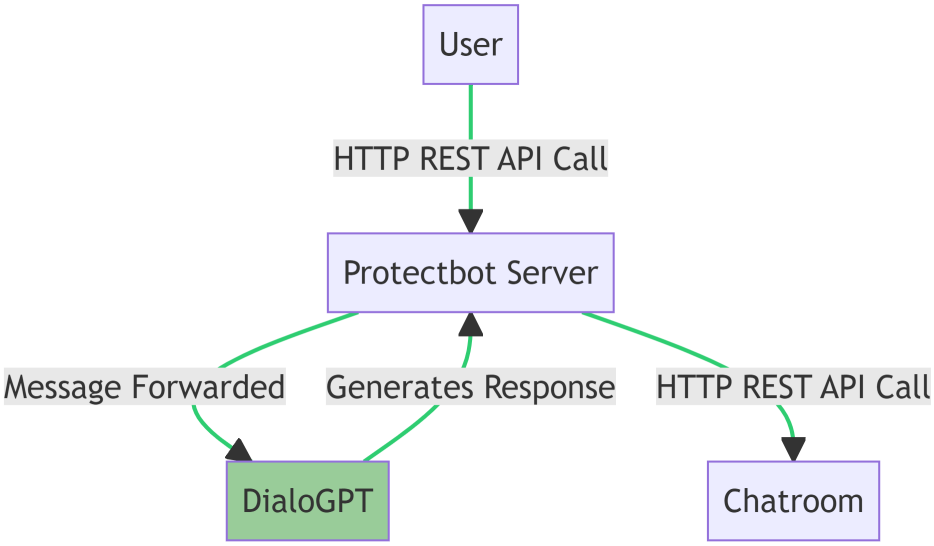

Figure 2. Protectbot conversational language generation model architecture (Steps 1–4).

The mathematical foundation of DialoGPT's attention mechanism, crucial for generating contextually relevant responses, is expressed as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q (query), K (key), and V (value) matrices represent the current input, conversation history, and potential responses, respectively. This formula enables the model to weigh the importance of different parts of the conversation, ensuring that responses are contextually appropriate. The normalization by $\sqrt{d_k}$ (the dimension of the keys) is a critical factor in scaling the dot products to manageable levels, thereby optimizing the softmax function's effectiveness.

By integrating a transformer-based mechanism, DialoGPT navigates the challenges of conversational context—marked by its informality, noise, and error-proneness—with finesse. The model's deep learning architecture, influenced by the original OpenAI GPT model [52], utilizes the attention mechanism to comprehend and generate responses, ensuring dynamic and context-aware engagement with users.

DialoGPT's conversational adeptness is further illustrated through its deployment in Protectbot [53], where it engages in real-time chat with potential suspects on gaming platforms. The adjustments made for conversational engagement, such as an extended vocabulary, increased context size, and optimized batch size, alongside training on a dataset of conversation-like exchanges from Reddit, equip DialoGPT with a nuanced understanding of dialogues [54].

Figures 3 and 4 depict the subsequent stages in Protectbot's operation, focusing on the classification of predatory behavior. After engaging in conversation, the chat logs are processed through a specialized classification model. This model employs a series of steps—preprocessing, feature extraction, and classification—to analyze the conversations and identify potential predatory behavior.

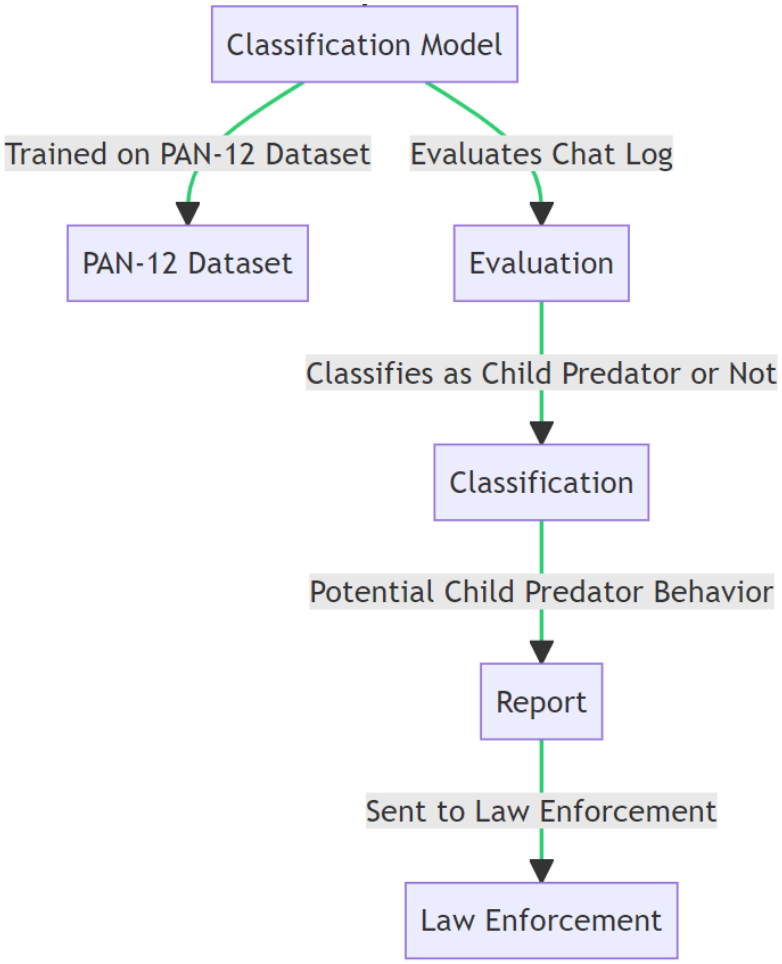

Figure 3. Architecture of Protectbot's classification model: sequential process from conversational engagement (Step 5) to behavioral analysis (Step 6).

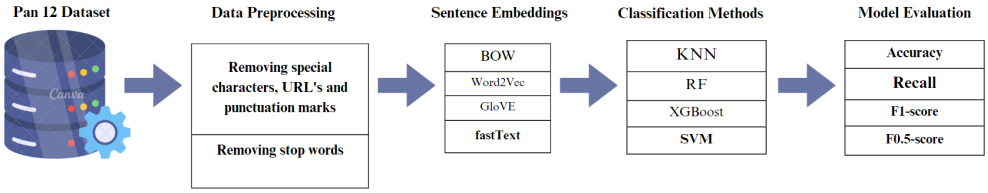

Figure 4. Comprehensive pipeline of the Protectbot classification model: process from initial data input through preprocessing, feature extraction, and final classification.

In summary, Protectbot's employment of DialoGPT, enriched with strategic modifications and an extensive training regimen, embodies a forward-thinking application of AI in the domain of child safety online. The attention-based transformer mechanism, in conjunction with a robust classification pipeline, sets a benchmark for AI-driven interventions against online predatory behaviors, ensuring a safer digital environment for vulnerable populations. From a user's perspective, interactions with Protectbot are designed to be indistinguishable from human conversations, ensuring users feel comfortable and engaged, thereby enhancing the system's effectiveness in identifying potential threats.

Adaptability and Scalability

Protectbot's architecture is designed for adaptability and scalability, enabling its deployment across different languages and cultural contexts with minimal adjustments. The conversational model can be fine-tuned with localized datasets to cater to specific linguistic nuances and slang, ensuring relevance and effectiveness in diverse online communities. This scalability extends Protectbot's potential applications beyond gaming platforms, offering a versatile tool in the fight against online predation in various digital spaces.

3.2. Machine Learning Classification Model

The Machine Learning Classification Model is intricately designed to scrutinize conversations generated by its conversational counterpart, aiming to discern whether interactions hint at predatory behavior towards children. This evaluative process is not arbitrary; it is meticulously structured around a dataset—specifically, the PAN12 dataset—comprising chat logs annotated based on their origins, either from child predators or non-predatory interactions. This dataset serves as the foundational training material, enabling the model to learn and distinguish between benign and harmful communications.

Operationally, the model is hosted on a server setup designed to seamlessly integrate with the conversational model through a PHP API, facilitating real-time analysis of chat logs. Upon receipt, each chat log undergoes a rigorous evaluation process. This is not merely a binary classification; it is an in-depth analysis aimed at identifying nuances and patterns indicative of predatory intent. When potential child predator behavior is identified, the system escalates its findings by compiling a detailed report—encompassing the chat log and pertinent details—for submission to law enforcement, thus enabling a timely and informed response.

The delineation of the classification model's pipeline offers a glimpse into the sophisticated architecture underlying its functionality. Initially, the model ingests raw text data—chat logs previously generated—serving as the input for the subsequent stages. Preprocessing is the first critical step, wherein the text is cleansed and standardized, removing any irrelevant or misleading information that could compromise the analysis.

Following preprocessing, the model employs feature extraction techniques, notably utilizing word embeddings to transform the textual data into a structured, analyzable format. These embeddings play a pivotal role, capturing the semantic essence of words and phrases within the conversations, thereby providing a rich, contextual basis for classification.

The selection of an optimal classification algorithm is paramount, as it directly influences the model's ability to accurately identify predatory behavior. This decision is predicated on a balance between computational efficiency, accuracy, and the algorithm's

suitability for text analysis. Once selected, the algorithm is trained on the preprocessed and structured dataset, fine-tuning its parameters for optimal performance.

The final stages of the pipeline are dedicated to evaluation, a bifurcated process involving both the prediction of outcomes on a test set and a comprehensive assessment of the model's overall performance. This evaluative phase is crucial for ensuring the model's efficacy and reliability in live scenarios, allowing for adjustments and refinements based on its predictive accuracy and the specificity of its classifications.

3.2.1. User Privacy and Data Security Measures

In developing Protectbot, paramount importance is placed on user privacy and data security. The system employs state-of-the-art encryption protocols for data in transit and at rest, adhering to global standards such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Regular audits and compliance checks ensure that Protectbot's operations respect user confidentiality and the integrity of collected data, fostering a secure online environment for all users.

3.2.2. Data Acquisition

The efficacy of our classification model hinges on the detailed analysis and selection of two critical datasets: the PAN12 dataset for training and a specially curated dataset from the PJ website for evaluation. These datasets are foundational to our goal of developing a model adept at discerning predatory behaviors in online conversations. This section delves into the specifics of these datasets, shedding light on their structure, content, and the pivotal roles they play in both the training and validation phases of our model.

PAN12 Dataset

Launched in 2012, the PAN12 dataset is a comprehensive compilation of textual data drawn from a wide array of genres, including, but not limited to, news articles, blogs, and social media posts. This diversity is crucial for training a model that is capable of accurately interpreting the nuances of various forms of communication. The dataset is divided into a training corpus and a test corpus for the purpose of detecting predatory behavior. It includes 66,927 conversations in the training set, with 2016 identified as predatory, juxtaposed against 64,911 non-predatory conversations. The test set expands this collection with 155,128 conversations, 3737 of which are classified as predatory. This imbalance is intentional, reflecting the real-world rarity of predatory interactions within the vast sea of online communication.

The PAN12 dataset is meticulously structured in XML format, ensuring detailed identification and analysis of each conversation through unique identifiers, timestamps, and author IDs. This structured approach is vital for nuanced analysis and effective model training. The dataset's complexity is further increased by the inclusion of adult, yet non-predatory, conversations, which poses a challenge in minimizing false positive rates.

Table 1 below provides a concise overview of the PAN12 dataset's composition, offering insight into its structure and the challenge posed by its imbalance.

Table 1. Overview of the PAN12 dataset structure.

Metric	Training Dataset	Test Dataset
Total Conversations	66,927	155,128
Predatory Conversations	2016	3737
Identified Predators	142	254

PJ Website Dataset

To rigorously evaluate our model, we sourced 71 complete predatory conversations from the PJ website, dated between 2013 and 2016, to avoid overlap with the PAN12 dataset. Compiled in CSV format, this dataset is crucial for assessing the real-world efficacy of our

model. It is publicly available, enhancing transparency and fostering further research in the field. The dataset [55] serves as a valuable resource for ongoing and future investigations into online safety and predatory behavior detection.

Figure 5 illustrates an example of a predatory conversation from the PAN12 dataset, highlighting the structured data format that includes detailed interaction attributes such as message sequence, author identity, timestamp, and textual content. This example showcases the dataset's capability to provide a detailed and structured framework for analysis.

```
<message line="4">
  <author>54b595f1920b5b1988e907ea693303b4</author>
  <time>00:02</time>
  <text>i guess its late, i wished i could've seen u</text>
</message>
<message line="5">
  <author>54b595f1920b5b1988e907ea693303b4</author>
  <time>00:02</time>
  <text>would've been fun</text>
</message>
<message line="6">
  <author>54b595f1920b5b1988e907ea693303b4</author>
  <time>00:02</time>
  <text>ok, still bored</text>
</message>
<message line="7">
  <author>54b595f1920b5b1988e907ea693303b4</author>
  <time>00:02</time>
  <text>hehe. i really gotta stop doin that</text>
</message>
<message line="8">
  <author>54b595f1920b5b1988e907ea693303b4</author>
  <time>00:02</time>
  <text>we could've had sex</text>
</message>
<message line="9">
  <author>54b595f1920b5b1988e907ea693303b4</author>
  <time>00:02</time>
  <text>kidding bout that</text>
</message>
```

Figure 5. Structured example from the PAN12 dataset, illustrating detailed interaction attributes.

The careful selection and analysis of these datasets underscore the meticulous approach adopted in developing a classification model that is both accurate and reliable in identifying potential child predators online. By integrating detailed descriptions and visual examples of the datasets, we aim to provide a comprehensive overview that underscores the methodological rigor of our research effort.

3.2.3. Text Cleaning and Prefiltering

The preparation of the PAN12 dataset for integration with the word embeddings model necessitated several preprocessing steps to ensure compatibility and enhance the dataset's utility for our analysis. Initially, the dataset's XML files were converted into a more manageable CSV format. This conversion was facilitated using Python's default XML parser, resulting in a streamlined dataset comprising four essential columns: conversation ID, author ID, message text, and conversation label. For the purposes of our analysis, conversations were dichotomously labeled: '0' for non-predatory and '1' for predatory. A conversation was classified as predatory if at least one participant was identified as a sexual predator. This is in line with the understanding that sexual predators predominantly engage in one-on-one interactions with their potential victims. Consequently, to reflect this behavioral pattern, conversations involving more than two participants, or solo author interactions, were excluded from the dataset [8,11,13,14,27,49,56].

Given the varying lengths of conversations within the PAN12 dataset, we recognized that those comprising very few messages posed a significant challenge for accurate classification. Specifically, conversations with fewer than six messages were deemed insufficient for reliable predatory or non-predatory determination. This decision to exclude shorter conversations was informed by the work of several researchers in the field [8,11,13,14,27,49,56], acknowledging the complexity and nuance required for effective predatory behavior analysis.

As a result of these prefiltering criteria, the dataset experienced a reduction in volume, culminating in a refined dataset. The training set was narrowed down to 8783 non-predatory and 973 predatory conversations, while the test dataset was adjusted to include 20,608 non-predatory and 1730 predatory conversations. This reduction is detailed in Table 2 below, which offers a succinct overview of the dataset's structure following the exclusion of short conversations.

Table 2. PAN12 dataset structure after prefiltering short conversations.

Number of Conversations	Training Dataset	Test Dataset
Non-predatory Conversations	8783	20,608
Predatory Conversations	973	1730

This meticulous approach to text cleaning and prefiltering was instrumental in ensuring that our dataset was optimally prepared for the subsequent stages of model training and evaluation. By focusing on conversations that provided sufficient contextual depth for analysis, we aimed to enhance the accuracy and reliability of our classification model in identifying predatory behavior online. The strategic reduction of the dataset underscores the importance of quality over quantity in the pursuit of meaningful and actionable insights.

3.2.4. Preprocessing

The preprocessing stage was pivotal in transforming the raw chat messages from the PAN12 dataset into a form amenable to semantic analysis and model training. Given the informal nature of online chats, these messages frequently eschew standard grammatical conventions, featuring a litany of misspellings, acronyms, and informal expressions, such as "*h r u?*" for "how are you?". Additionally, the text often includes elongated words for emphasis, like "*heyyyyy*" or "*soryyyyy*", which, while expressive, present challenges for text processing.

In this crucial phase of data preparation, several modifications were applied to the dataset to enhance its suitability for NLP tasks:

- Removal of Non-Textual Elements: URLs, punctuation marks, and special characters, including symbols like $, &, #, +, and =, were systematically removed from the dataset. Numerical values, which are often irrelevant to the context of predatory behavior analysis, were also excluded.
- Exclusion of Stop Words: Commonly occurring stop words, including but not limited to *is*, *are*, *the*, and *a*, were removed from the dataset. These words, while prevalent in natural language, typically do not contribute to the semantic richness required for text mining and were thus omitted to streamline the dataset for more efficient processing.

While some practitioners advocate for the elimination of misspelled words and the standardization of chat messages to conform to conventional spelling and grammar, such an approach was deliberately avoided in our preprocessing strategy. This decision was informed by the recognition that misspellings and informal expressions can carry significant behavioral indicators, particularly in the context of predatory behavior [8,11,13,14,27,49,56]. The elimination of these elements could inadvertently strip the dataset of nuanced information critical for identifying predatory patterns.

- Retention of Original Textual Features: Notably, our preprocessing did not involve stemming or lemmatization processes. These techniques, while useful for reducing words to their root forms, were omitted to preserve the integrity and informational value of the original text. The decision to forgo these steps was made to ensure that the dataset retained the maximum possible amount of relevant information, acknowledging that the unique linguistic characteristics present in chat messages might hold key insights into predatory behaviors.

This preprocessing approach underscores our commitment to maintaining the dataset's richness and complexity, recognizing that the idiosyncratic features of online communication can be invaluable in distinguishing between predatory and non-predatory interactions. By carefully balancing the need for clean, analyzable data with the desire to preserve the dataset's intrinsic informational value, we aimed to enhance the accuracy and sensitivity of our classification model in detecting potential threats within online conversations.

3.2.5. Word Embeddings and Classification

Representing textual data numerically is pivotal in ML for facilitating the processing and analysis of natural language. Unlike images or audio, text requires a transformation into high-dimensional vector spaces to encode its semantic content effectively. This transformation allows sentences to be interpreted as points within a multidimensional semantic space, where proximity denotes similarity in meaning. Such a representation is crucial for tasks like sentence similarity assessments and text classification.

Generating Sentence Vectors

The construction of sentence vectors often begins with the representation of individual words as vectors. These word vectors are then aggregated to form a sentence vector that captures the overall semantic essence of the sentence. Among the myriad techniques for creating word vectors, two prominent methods stand out: the Bag of Words (BoW) model and word embeddings:

- Bag of Words (BoW): Utilized extensively for generating sentence vectors, BoW creates a vocabulary from a corpus and assigns an index to each word in a high-dimensional vector space. The presence of words within a text is marked by non-zero entries in the corresponding vector, with values representing binary occurrences, term counts, or term frequency-inverse document frequency (TF-IDF) scores. However, BoW's limitations are its inability to capture semantic relationships between words and the impracticality of managing large, sparse vectors for extensive corpora.
- Word Embeddings: This approach overcomes BoW's shortcomings by mapping words with similar meanings to proximate points in the vector space, thus preserving semantic relationships. Popular methods include Word2Vec [57], GloVe [58], and fastText [59], with each employing distinct mechanisms for generating word vectors. Word2Vec and GloVe rely on contextual relationships or co-occurrence matrices, while fastText, innovatively, uses character n-grams to accommodate out-of-vocabulary words. This characteristic enables fastText to generate meaningful vectors for previously unseen words, showcasing its versatility and superior performance in various comparative studies [60,61].

Application to Predatory Behavior Detection

Our approach to detecting predatory behavior within the PAN12 dataset leveraged fastText's "cc.en.300" pre-trained model to transform chat messages into 300-dimensional sentence vectors. The "cc.en.300" model, trained on an extensive corpus comprising over 600 billion words from diverse sources, offers a rich linguistic foundation for accurately capturing the semantic nuances of chat logs [59]. These sentence vectors served as the input for training several classifiers, including K-nearest neighbors (KNN), support vector machines (SVMs), Random Forest (RF), and XGBoost, each evaluated for its efficacy in distinguishing predatory conversations.

The performance assessment of these classifiers was conducted not only on the PAN12 test dataset, but also on a curated dataset derived from the PJ website, ensuring a comprehensive evaluation of the model's ability to identify predatory behavior accurately. This dual-dataset strategy underscores our commitment to validating the model's robustness and reliability across different sources of textual data, highlighting the effectiveness of fastText embeddings coupled with advanced classification techniques in safeguarding online interactions. The model's efficacy in identifying predatory behavior was validated through rigorous testing with real-world data, demonstrating its potential to significantly enhance online safety measures for children. In developing Protectbot, ethical considerations, particularly regarding privacy and data protection, were paramount. Measures to anonymize data and ensure its secure handling reflect our commitment to responsible AI use.

3.3. System Integration

The Protectbot framework's efficacy hinges on the seamless integration between its conversational language generation and classification models. This integration is facilitated through a web-based interface utilizing HTTP REST API calls, enabling dynamic data exchange between the two core components of the system. These calls can be considered as a way for two computer systems to communicate over the internet using standard web protocols. This process ensures secure and efficient data exchange, crucial for real-time analysis and response to detected predatory behavior. This subsection delves into the mechanisms of this integration, illustrating how the Protectbot system manages the flow of information from initial conversation to potential predator identification.

3.3.1. Real-World Deployment and Feedback

The deployment of Protectbot in online gaming environments has yielded valuable insights into its operational effectiveness and user engagement. Collaborative efforts with gaming platforms and law enforcement agencies have facilitated several pilot tests, resulting in actionable reports that led to preventive measures against potential predatory activities. Feedback from these real-world applications has been instrumental in refining the system's response algorithms and enhancing its conversational authenticity, ensuring a more proactive and nuanced approach to online child safety.

Interface for Model Communication

The cornerstone of Protectbot's system integration is the communication interface established between the conversational and classification models. This interface operates through HTTP REST API calls, a standard web connection protocol that allows for the efficient transmission of chat logs and classification results between the models. Upon the completion of a user interaction, the conversational model transmits the compiled chat log to the classification model via this dedicated web connection.

Classification and Response

Upon receiving the chat log, the classification model embarks on a critical process of analysis. It evaluates the textual content, applying its trained algorithms to discern patterns indicative of predatory behavior. This analysis culminates in a binary classification, categorizing the conversation as either "potential predator" or "not a potential predator". The outcome of this classification is then communicated back to the conversational model, informing the next steps in the Protectbot system's response protocol.

Actionable Outcomes

The classification result triggers a decision-making process within the Protectbot system. If the classification model identifies the conversation as indicative of potential predator behavior, the system initiates the generation of a detailed report. This report encompasses the chat log in question along with the classification outcome, providing a

comprehensive overview for further review. The prepared report is then dispatched to relevant law enforcement agencies, offering them a foundation for potential investigative actions. This proactive approach underscores the Protectbot system's commitment to leveraging technological innovation for enhancing online safety and security.

Visualizing System Integration

The entirety of the Protectbot system's operation, from the initiation of chat to the final decision on potential predator behavior, is encapsulated in the integration process depicted in Figure 6. This visual representation outlines the complete six-step workflow, highlighting the critical role of system integration in facilitating the transition from conversational engagement to actionable intelligence. The diagram serves as a succinct overview of the Protectbot framework's operational dynamics, emphasizing the synergy between conversational interaction and behavioral classification.

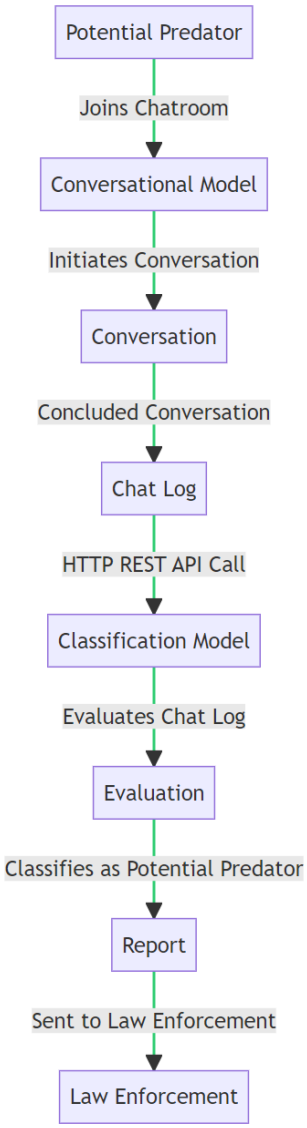

Figure 6. Overview of Protectbot system integration: from initial chat to potential predator identification.

In summary, the integration of the conversational language generation model with the classification model forms the backbone of the Protectbot system, enabling it to identify potential predatory behavior within online conversations effectively. This integrated approach not only exemplifies the application of advanced AI and ML techniques in real-

world scenarios, but also showcases the potential of technology-driven solutions to address pressing societal concerns.

4. Results

This section provides a comparative analysis of machine learning classifiers—K-Nearest Neighbors (KNN), SVM, RF, and XGBoost—utilizing the PAN12 dataset enhanced with pre-trained fastText embeddings. Our binary classification framework distinguishes between non-predatory (Class 0) and predatory conversations (Class 1), evaluating performance based on accuracy, recall, F_1-score, and $F_{0.5}$-score.

4.1. Classification Performance

Table 3 encapsulates the performance metrics for each classifier. It illustrates the classifiers' abilities to distinguish between predatory and non-predatory conversations based on the PAN12 dataset. The metrics considered for evaluation include accuracy, recall, F_1-score, and $F_{0.5}$-score for both classes individually and as a weighted average across the dataset.

Table 3. Performance metrics for ML classifiers on the PAN12 dataset.

Classifier	Class 0 (Non-Predatory)				Class 1 (Predatory)				Weighted Avg			
	Accuracy	Recall	F_1-Score	$F_{0.5}$-Score	Accuracy	Recall	F_1-Score	$F_{0.5}$-Score	Accuracy	Recall	F_1-Score	$F_{0.5}$-Score
XGBoost	0.99	1	0.99	0.98	0.96	0.84	0.90	0.93	0.98	0.92	0.95	0.95
KNN (n = 15, p = 2)	1	0.98	0.99	0.99	0.83	0.95	0.88	0.85	0.92	0.97	0.94	0.92
SVM (C = 10, γ = 1, kernel = rbf)	0.99	1	1	0.99	0.96	0.93	0.95	0.95	0.98	0.97	0.98	0.97
RF	0.97	1	0.99	0.98	0.98	0.69	0.81	0.90	0.98	0.85	0.90	0.94

The SVM classifier emerges as the most effective, showing superior performance across all metrics. This suggests that SVM's algorithmic strengths in handling high-dimensional data and its ability to model complex nonlinear relationships make it particularly suited for this classification task. However, the KNN classifier's high recall rate for Class 1 indicates its potential usefulness in scenarios where identifying every possible instance of predatory behavior is critical, despite the trade-off in precision.

4.2. Loss Metrics Evaluation

The evaluation of Mean Absolute Error (MAE) and Mean Squared Logarithmic Error (MSLE) provides an in-depth look at the performance of classifiers during both training and testing phases. This analysis is crucial for understanding the models' generalization capabilities and identifying any overfitting issues. Table 4 showcases these metrics for each classifier, allowing for a direct comparison of their robustness and reliability.

Table 4. Loss metrics for ML classifiers on the PAN12 dataset.

Classifier	Training Set		Testing Set	
	MAE	MSLE	MAE	MSLE
XGBoost	0	0	0.015	0.007
KNN	0.023	0.011	0.019	0.009
SVM	0.008	0.004	0.008	0.003
RF	0	0	0.024	0.011

The examination of MAE and MSLE values across the training and testing phases highlights the classifiers' varying degrees of overfitting and generalization. XGBoost and Random Forest (RF) exhibit a perfect performance during the training phase with zero error, which, unfortunately, does not carry over to the testing phase, indicating a high propensity for overfitting. On the contrary, the support vector machine (SVM) classifier demonstrates remarkable consistency between training and testing errors, specifically maintaining the

lowest testing errors among all classifiers. This underscores SVM's superior capability to generalize across different data samples, making it a more robust and reliable model for practical applications where the ability to perform well on unseen data is critical.

4.3. Comparative Analysis with State of the Art

This subsection conducts a comparative analysis of our model against state-of-the-art text classification techniques, utilizing the PAN12 dataset. The performance of our proposed SVM classifier, augmented with pre-trained FastText embeddings, is benchmarked against various methodologies reported in recent literature. Table 5 highlights this comparison across multiple performance metrics.

Table 5. Comparative analysis of text classification models utilizing the PAN12 dataset.

Technique	Paper	Precision	Accuracy	Recall	F_1-Score	$F_{0.5}$-Score
BoW with TF-IDF Weighting + NN	[62]	0.98	-	0.78	0.87	0.93
BoW with TF-IDF Weighting + Linear SVM	[11]	-	0.98	-	-	-
Soft Voting: BoW with TF-IDF Weighting + SVM, BoW with Binary Weighting + MNB, BoW with Binary Weighting + LR	[14]	1.0	0.99	0.95	0.98	0.99
Word2Vec + CNN	[46]	0.29	0.88	0.70	0.42	-
Word2Vec + Class Imbalance + Histogram Gradient Boosted DT	[12]	-	0.99	-	0.99	0.94
BoW + SVM + RF	[63]	1	-	0.82	-	0.957
CNN + Multilayer Perceptron	[49]	0.46	-	0.72	-	-
BERT + Feed Forward NN	[8]	0.98	-	0.99	0.98	0.98
TF-IDF + SVM	[56]	0.92	0.91	0.89	0.91	0.91
One-hot CNN	[64]	0.92	-	0.72	0.81	-
Pre-Trained FastText + SVM (Ours)	This Study	0.99	0.99	0.99	0.99	0.99

Table 5 delineates a meticulous comparative analysis, placing our methodology at the pinnacle in terms of precision, accuracy, recall, F_1-Score, and $F_{0.5}$-Score, with each metric peaking at 0.99. This not only demonstrates the effectiveness of the pre-trained FastText with SVM approach but also its unprecedented achievement in the realm of predatory chat detection. Our model's uniform excellence across these metrics signifies its capability to balance both precision and recall effectively, minimizing false positives and false negatives in a domain where accuracy is paramount. The comparison with state-of-the-art models, ranging from traditional BoW techniques to sophisticated neural network architectures like BERT, establishes our model's supremacy in dealing with complex text classification challenges. This analysis reinforces our model's position as a groundbreaking tool in the ongoing effort to enhance online safety and security.

4.4. Overfitting Assessment and Model Generalization

To ensure the robustness and generalizability of our proposed model, we conducted a thorough overfitting assessment using a 10-fold cross-validation technique on the training dataset. This method provides a more reliable estimate of model performance on unseen data by partitioning the dataset into ten subsets, training the model on nine subsets, and validating it on the remaining subset. This process is repeated ten times, with each subset serving as the validation set once.

The cross-validation results indicate that our model maintained consistent performance across all folds, suggesting a strong generalization capability. The accuracy and

recall metrics obtained from the cross-validation closely align with those observed on the test dataset, underscoring the model's ability to generalize well to new, unseen data.

This consistency in performance across different subsets of data is indicative of the model's resilience against overfitting, a common challenge in machine learning models where a model performs well on training data, but poorly on new data. The successful mitigation of overfitting enhances the model's applicability in real-world scenarios, where it is crucial for models to perform accurately on data not encountered during the training phase.

4.5. Confusion Matrix Analysis

The confusion matrix provides a comprehensive visual and quantitative analysis of a classifier's performance, highlighting the distribution of true positives, false positives, false negatives, and true negatives. This study delves into the confusion matrices of the SVM, KNN, RBF, and XGBoost classifiers when applied to the PAN12 dataset, offering a nuanced view of each model's ability to distinguish between non-predatory and predatory conversations. Figure 7 reveals the intricate details of each classifier's performance.

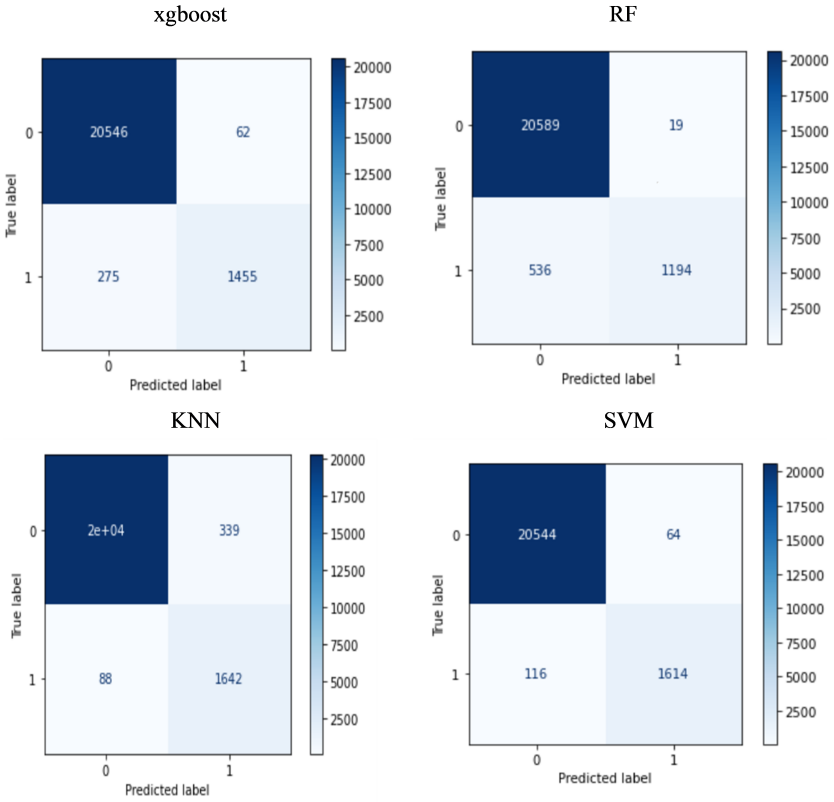

Figure 7. Detailed Confusion matrix visualization for classifiers.

Analysis of Figure 7 reveals the specific strengths and weaknesses of the classifiers. The SVM classifier stands out for its precision, accurately classifying 20,544 out of 20,608 non-predatory conversations and correctly identifying 1614 out of 1730 predatory chats. This high level of accuracy is crucial for applications where minimizing false positives is essential, such as in monitoring systems where unnecessary alerts could dilute the focus on genuine threats.

Conversely, the KNN classifier demonstrates a higher tendency towards false negatives but fewer false positives, indicating its effectiveness in scenarios where failing to detect a positive instance (a potential threat) is more detrimental than incorrectly labeling a negative instance (a safe interaction). This characteristic makes the KNN classifier a valuable option for platforms that prioritize the detection of predatory behavior, despite the risk of overlooking some non-threatening interactions.

The detailed examination provided by the confusion matrices not only underscores the individual strengths and limitations of each classifier, but also assists in selecting the most suitable model based on the specific needs of a deployment environment. Future research could explore the optimization of classifier parameters and the potential of ensemble methods to further enhance accuracy and balance the trade-off between false positives and false negatives.

4.6. Validation on Curated Dataset

The validation of our model on a curated dataset further underscores its effectiveness in identifying predatory behavior within chat logs. This curated dataset, distinct from the PAN12 dataset, provides a real-world testing ground to assess the practical applicability of our classifiers, specifically the SVM and KNN models, which have shown promising results in earlier evaluations.

The ensuing results are tabulated in Table 6, which underscores the effectiveness of our approach in discerning predatory behaviors within chat logs. This evaluation reveals the nuanced performance of the KNN and SVM classifiers, highlighting their strengths and potential limitations in a real-world context.

Table 6. Performance evaluation of KNN and SVM classifiers on the PJ-based curated dataset, highlighting true positive and false negative rates.

Classifier	Correctly Classified	Incorrectly Classified	TP Rate	FN Rate
KNN	66	5	0.92	0.08
SVM	48	23	0.68	0.32

The KNN classifier exhibits a high true positive rate (TP Rate) of 0.92, correctly identifying 66 out of 71 predatory chats, with only a minimal number of chats misclassified. This demonstrates the KNN model's sensitivity and its potential as a robust tool for online safety applications, where accurately detecting predatory behavior is crucial.

Conversely, the SVM classifier, while showing a lower true positive rate of 0.68, correctly classified 48 predatory chats. The higher false negative rate (FN Rate) of 0.32 highlights some challenges in accurately identifying all predatory instances. However, its considerable success in correctly identifying a significant portion of predatory chats underscores its utility in scenarios where a balance between precision and recall is desired.

The validation on this curated dataset not only confirms the models' capabilities in detecting predatory behavior, but also highlights their distinct strengths and limitations when applied to different data types. It reinforces the need for a nuanced approach to classifier selection, tailored to the specific objectives and constraints of each application. Further research could explore adaptive models or hybrid approaches that combine the strengths of SVM and KNN classifiers to achieve even higher accuracy and reliability in real-world settings.

4.7. Discussion

The assessment of our model across both the PAN12 and specially curated datasets has solidified its capability in effectively detecting predatory behavior within chat logs. This section revisits and expands upon the contributions mentioned in the Introduction, providing a more comprehensive discussion of our results in relation to these key areas:

- Advanced Text Classification for Online Child Safety: Our SVM classifier's exceptional performance underscores a significant leap in text classification methodologies aimed at online child safety. Achieving metrics of accuracy, precision, recall, F_1-score, and $F_{0.5}$-score at approximately 0.99 for all, the SVM classifier not only validates the efficacy of our text classifier in identifying predatory behavior, but also positions our approach as a leading methodology within the realm of AI-based child protection. This aligns with our contribution towards enhancing child safety through advanced AI techniques.

- Nuanced Classifier Selection Based on Application Needs: Highlighting the KNN classifier's superior recall rate for predatory conversations points to its effectiveness in scenarios where the cost of missing a potential threat is high. This finding supports our contribution of presenting a nuanced, application-specific approach to classifier selection, emphasizing that platforms prioritizing the reduction of false negatives might find KNN more suitable, despite its slightly lower overall accuracy compared to SVM.
- Insights from Loss Metrics on Model Generalization: Our detailed analysis of loss metrics, specifically Mean Absolute Error (MAE) and Mean Squared Logarithmic Error (MSLE), provides critical insights into the models' ability to generalize. SVM's consistently low error rates across both training and testing phases highlight its robustness and reliability, crucial for deployment in unpredictable real-world settings. This supports our contribution of refining text classification strategies for child safety, showcasing the importance of model generalization in practical applications.
- Comparative Analysis with State-of-the-Art Models: The combination of pre-trained FastText embeddings with the SVM classifier demonstrates significant advantages in detecting predatory behavior, marking a notable advancement in the field. This performance not only affirms the effectiveness of our methods, but also paves the way for future explorations into enhancing AI-driven child safety solutions. Our contribution in this area highlights the potential for innovative embedding techniques combined with classical ML models to improve online child protection.
- Challenges and Future Directions in Predator Detection: The validation process on the curated dataset illuminated the inherent challenges in accurately identifying predatory behavior, revealing the complex nature of this task. The variable performance of SVM and KNN on this dataset points to the potential benefits of employing hybrid models or adaptive techniques to enhance accuracy and reliability in predator detection. This discussion emphasizes our commitment to future research aimed at overcoming these challenges, aligning with our initial contribution towards advancing the field of AI in child online safety.

5. Conclusions and Future Work

We developed a comprehensive framework, Protectbot, aimed at mitigating predatory behavior on various gaming platforms. Our primary innovation lies in the advanced text classifier that employs state-of-the-art techniques, including the use of fastText for generating word embeddings and the application of ML classifiers such as SVM, KNN, RF, and XGBoost. The classifier, rigorously trained and evaluated on the PAN12 dataset, demonstrated exceptional performance, notably with the SVM classifier, achieving metrics of accuracy, precision, recall, F_1-score, and $F_{0.5}$-score at 0.99 for all. These results not only validate the efficacy of our text classifier in identifying predatory behavior within gaming chat platforms, but also position our approach at the forefront of current text classification methodologies.

Furthermore, the validation of our classifiers on a newly curated dataset, comprising real-world predatory chats, underscores the robustness and reliability of the SVM classifier and highlights the KNN's potential in accurately detecting predatory behavior. These findings affirm the significant potential of Protectbot in enhancing the safety of online gaming communities.

As we look toward the future, our research opens up multiple pathways for advancement and implementation. The direct integration of the Protectbot framework into gaming platforms presents a challenge that necessitates a thoughtful examination of scalability, user experience, and the technical feasibility of embedding Protectbot as an interactive entity within the gaming environment. Achieving this integration would not only enhance the framework's ability to provide real-time protection but also ensure a seamless user experience without disrupting the gaming ecosystem.

To further refine Protectbot's conversational capabilities, we propose leveraging Large Language Models (LLMs) to enrich the chatbot's responses, making them more reflective of children's linguistic patterns. This approach, utilizing incoming messages as prompts for a finely tuned model such as DialoGPT or an advanced LLM, is anticipated to significantly improve the chatbot's realism and its capacity to engage predators effectively.

Navigating the ethical, privacy, and legal landscapes remains paramount in our journey toward realizing these objectives. Prioritizing data anonymity, securing informed consent, optimizing threat detection accuracy, and addressing model biases are crucial for ethical deployment. Collaborative efforts with law enforcement, adherence to legal frameworks, system security against unauthorized access, and continuous evaluation will underpin a responsible and impactful application.

Our future work will extend beyond the practicalities of embedding Protectbot within live gaming contexts to embrace the broader challenges of deploying AI in child protection. This includes a collaborative approach with game developers to ensure respectful integration into the gaming experience while providing robust protection. Furthermore, we aim to conduct a more detailed comparative analysis with other AI-based child protection solutions, exploring their features, detection capabilities, and user experience to provide a comprehensive understanding of Protectbot's contributions to the field. Additionally, discussions on potential enhancements, such as incorporating real-time adaptive learning or exploring multimodal data analysis, will offer a roadmap for advancing AI's role in child online safety.

In conclusion, Protectbot contributes significantly to safeguarding vulnerable online communities, particularly within the gaming environment. By outlining a path for future research and development, we inspire further innovations in this critical area, ensuring that digital spaces remain safe and inclusive for all users.

Author Contributions: Conceptualization, J.M. and A.F.; methodology, J.M. and A.F.; software, A.F. and F.A.; validation, A.F. and F.A.; formal analysis, J.M. and A.F.; resources, J.M. and A.F.; data curation, A.F. and F.A.; writing—original draft preparation, A.F.; writing—review and editing, J.M., I.K. and A.K.; visualization, J.M. and I.K.; supervision, J.M.; funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Academic Research Committee of Rochester Institute of Technology, Dubai.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The curated PJ dataset that supports the findings of this study is available at https://doi.org/10.21227/4kyv-n442, accessed on 4 March 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. American Psychological Association Resolution on Violent Video Games. Available online: http://www.apa.org/about/policy/violent-video-games.aspx (accessed on 4 March 2024).
2. Faraz, A.; Mounsef, J.; Raza, A.; Willis, S. Child Safety and Protection in the Online Gaming Ecosystem. *IEEE Access* **2022**, *10*, 115895–115913. [CrossRef]
3. Digital 2021: Global Overview Report. Available online: https://datareportal.com/reports/digital-2021-global-overview-report (accessed on 4 March 2024).
4. Stalker, P.; Livingstone, S.; Kardefelt-Winthe, D.; Saeed, M. *Growing up in a Connected World*; UNICEF Office of Research–Innocenti: Firenze, Italiy, 2019.
5. Child Rights and Online Gaming: Opportunities & Challenges for Children and the Industry. Available online: https://www.unicef-irc.org/files/upload/documents/UNICEF_CRBDigitalWorldSeriesOnline_Gaming.pdf (accessed on 4 March 2024).
6. Helbing, D.; Brockmann, D.; Chadefaux, T.; Donnay, K.; Blanke, U.; Woolley-Meza, O.; Moussaid, M.; Johansson, A.; Krause, J.; Schutte, S.; et al. Saving Human Lives: What Complexity Science and Information Systems can Contribute. *J. Stat. Phys.* **2015**, *158*, 735–781. [CrossRef]
7. Perc, M.; Ozer, M.; Hojnik, J. Social and Juristic Challenges of Artificial Intelligence. *Palgrave Commun.* **2019**, *5*, 61. [CrossRef]

8. Agarwal, N.; Ünlü, T.; Wani, M.A.; Bours, P. Predatory Conversation Detection Using Transfer Learning Approach. In Proceedings of the 7th International Conference on Machine Learning, Optimization, and Data Science (LOD), Grasmere, UK, 4–8 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; Volume 13163, pp. 488–499.
9. Anderson, P.; Zuo, Z.; Yang, L.; Qu, Y. An Intelligent Online Grooming Detection System Using AI Technologies. In Proceedings of the International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23–26 June 2019; IEEE: New York, NY, USA, 2019; pp. 1–6.
10. Andleeb, S.; Ahmed, R.; Ahmed, Z.; Kanwal, M. Identification and Classification of Cybercrimes using Text Mining Technique. In Proceedings of the International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 16–18 December 2019; IEEE: New York, NY, USA, 2019; pp. 227–232.
11. Borj, P.R.; Bours, P. Predatory Conversation Detection. In Proceedings of the International Conference on Cyber Security for Emerging Technologies (CSET), Doha, Qatar, 27–29 October 2019; IEEE: New York, NY, USA, 2019; pp. 1–6.
12. Borj, P.R.; Raja, K.B.; Bours, P. Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles Effects. In Proceedings of the 20th International Conference of the Biometrics Special Interest Group (BIOSIG), Digital Conference, 15–17 September 2021; Gesellschaft für Informatik e.V.: Hamburg, Germany, 2021; Volume P-315, pp. 245–252.
13. Bours, P.; Kulsrud, H. Detection of Cyber Grooming in Online Conversation. In Proceedings of the International Workshop on Information Forensics and Security (WIFS), Delft, The Netherlands, 9–12 December 2019; IEEE: New York, NY, USA, 2019; pp. 1–6.
14. Fauzi, M.A.; Bours, P. Ensemble Method for Sexual Predators Identification in Online Chats. In Proceedings of the 8th International Workshop on Biometrics and Forensics (IWBF), Porto, Portugal, 29–30 April 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
15. Gunawan, F.E.; Ashianti, L.; Sekishita, N. A Simple Classifier for Detecting Online Child Grooming Conversation. *Telkomnika (Telecommun. Comput. Electron. Control)* **2018**, *16*, 1239–1248. [CrossRef]
16. Kick Ass Open Web Technologies IRC Logs. Available online: https://krijnhoetmer.nl/irc-logs/ (accessed on 4 March 2024).
17. Kim, J.; Kim, Y.J.; Behzadi, M.; Harris, I.G. Analysis of Online Conversations to Detect Cyberpredators Using Recurrent Neural Networks. In Proceedings of the 1st International Workshop on Social Threats in Online Conversations: Understanding and Management (STOC@LREC), Marseille, France, 12–13 May 2020; European Language Resources Association: Paris, France, 2020; pp. 15–20.
18. Kirupalini, S.; Baskar, A.; Ramesh, A.; Rengarajan, G.; Gowri, S.; Swetha, S.; Sangeetha, D. Prevention of Emotional Entrapment of Children on Social Media. In Proceedings of the International Conference on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India, 25–27 August 2021; IEEE: New York, NY, USA, 2021; pp. 95–100.
19. Laorden, C.; Galán-García, P.; Santos, I.; Sanz, B.; Hidalgo, J.M.G.; Bringas, P.G. Negobot: A Conversational Agent Based on Game Theory for the Detection of Paedophile Behaviour. In Proceedings of the International Joint Conference CISIS'12-ICEUTE'12-SOCO'12, Ostrava, Czech Republic, 5–7 September 2012; Springer: Berlin/Heidelberg, Germany, 2012; Volume 189, pp. 261–270.
20. Ngejane, C.H.; Eloff, J.H.P.; Sefara, T.J.; Marivate, V.N. Digital Forensics Supported by Machine Learning for the Detection of Online Sexual Predatory Chats. *Forensic Sci. Int. Digit. Investig.* **2021**, *36*, 301109. [CrossRef]
21. Pardo, F.M.R.; Rosso, P.; Koppel, M.; Stamatatos, E.; Inches, G. Overview of the Author Profiling Task at PAN 2013. In Proceedings of the Working Notes for CLEF Conference, CEUR-WS.org, Valencia, Spain, 23–26 September 2013; Volume 1179.
22. Perverted Justice Foundation. Available online: http://www.perverted-justice.com/ (accessed on 4 March 2024).
23. Ringenberg, T.R.; Misra, K.; Rayz, J.T. Not So Cute but Fuzzy: Estimating Risk of Sexual Predation in Online Conversations. In Proceedings of the International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; IEEE: New York, NY, USA, 2019; pp. 2946–2951.
24. Rodríguez, J.I.; Durán, S.R.; Díaz-López, D.; Pastor-Galindo, J.; Mármol, F.G. C^3-Sex: A Conversational Agent to Detect Online Sex Offenders. *Electronics* **2020**, *9*, 1779. [CrossRef]
25. Sulaiman, N.R.; Siraj, M.M. Classification of Online Grooming on Chat Logs Using Two Term Weighting Schemes. *Int. J. Innov. Comput.* **2019**, *9*. [CrossRef]
26. Triviño, J.M.; Rodríguez, S.M.; López, D.O.D.; Mármol, F.G. C3-Sex: A Chatbot to Chase Cyber Perverts. In Proceedings of the International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 5–8 August 2019; IEEE: New York, NY, USA, 2019; pp. 50–57.
27. Wani, M.A.; Agarwal, N.; Bours, P. Sexual-predator Detection System based on Social Behavior Biometric (SSB) Features. In Proceedings of the 5th International Conference on Arabic Computational Linguistics (ACLING), Virtual Event, 4–5 June 2021; Procedia Computer Science; Volume 189, pp. 116–127.
28. Zuo, Z.; Li, J.; Anderson, P.; Yang, L.; Naik, N. Grooming Detection using Fuzzy-Rough Feature Selection and Text Classification. In Proceedings of the International Conference on Fuzzy Systems (FUZZ-IEEE), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE: New York, NY, USA, 2018; pp. 1–8.

29. Zuo, Z.; Li, J.; Wei, B.; Yang, L.; Chao, F.; Naik, N. Adaptive Activation Function Generation for Artificial Neural Networks through Fuzzy Inference with Application in Grooming Text Categorisation. In Proceedings of the International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23–26 June 2019; IEEE: New York, NY, USA, 2019; pp. 1–6.
30. Inches, G.; Crestani, F. Overview of the International Sexual Predator Identification Competition at PAN-2012. In Proceedings of the CLEF 2012 Evaluation Labs and Workshop, CEUR-WS.org, Rome, Italy, 17–20 September 2012; CEUR Workshop Proceedings; Volume 1178.
31. Verma, K.; Davis, B.; Milosevic, T. Examining the Effectiveness of Artificial Intelligence-Based Cyberbullying Moderation on Online Platforms: Transparency Implications. *AoIR Sel. Pap. Internet Res.* **2022**. [CrossRef]
32. Halder, D. PUBG Ban and Issues of Online Child Safety during COVID-19 Lockdown in India: A Critical Review from the Indian Information Technology Act Perspectives. *Temida* **2021**, *24*, 303–327. [CrossRef]
33. Rita, M.N.; Shava, F.B.; Chitauro, M. Tech4Good: Artificial Intelligence Powered Chatbots with Child Online Protection in Mind. *Inf. Syst. Emerg. Technol.* **2022**, 35. https://www.researchgate.net/profile/Abubakar-Saidu-Arah-Phd/publication/372992925_Information_and_Communication_Technologies_Readiness_and_Acceptance_among_Teachers_in_Vocational_Enterprises_Institutions_in_Abuja_Nigeria/links/64d37471b684851d3d92fcbd/Information-and-Communication-Technologies-Readiness-and-Acceptance-among-Teachers-in-Vocational-Enterprises-Institutions-in-Abuja-Nigeria.pdf#page=49 (accessed on 4 March 2024).
34. Mohasseb, A.; Bader-El-Den, M.; Kanavos, A.; Cocea, M. Web Queries Classification Based on the Syntactical Patterns of Search Types. In Proceedings of the 19th International Conference on Speech and Computer (SPECOM), Hatfield, UK, 12–16 September 2017; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10458, pp. 809–819.
35. Mohasseb, A.; Kanavos, A. Grammar-Based Question Classification Using Ensemble Learning Algorithms. In Proceedings of the 18th International Conference on Web Information Systems and Technologies (WEBIST), Valletta, Malta, 25–27 October 2022; Lecture Notes in Business Information Processing; Springer: Berlin/Heidelberg, Germany, 2022; Volume 494, pp. 84–97.
36. Zambrano, P.; Sánchez, M.; Torres, J.; Fuertes, W. BotHook: An Option against Cyberpedophilia. In Proceedings of the 1st Cyber Security in Networking Conference (CSNet), Janeiro, Brazil, 18–20 October 2017; IEEE: New York, NY, USA, 2017; pp. 1–3.
37. Urbas, G. Legal Considerations in the Use of Artificial Intelligence in the Investigation of Online Child Exploitation. In *ANU College of Law Research Paper*; 2021. Available online: https://ssrn.com/abstract=3978325 (accessed on 4 March 2024).
38. Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani, A. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access* **2019**, *7*, 70701–70718. [CrossRef]
39. Fire, M.; Goldschmidt, R.; Elovici, Y. Online Social Networks: Threats and Solutions. *IEEE Commun. Surv. Tutor.* **2014**, *16*, 2019–2036. [CrossRef]
40. Jevremovic, A.; Veinovic, M.; Cabarkapa, M.; Krstic, M.; Chorbev, I.; Dimitrovski, I.; Garcia, N.; Pombo, N.; Stojmenovic, M. Keeping Children Safe Online With Limited Resources: Analyzing What is Seen and Heard. *IEEE Access* **2021**, *9*, 132723–132732. [CrossRef]
41. de Morentin, J.I.M.; Lareki, A.; Altuna, J. Risks Associated with Posting Content on the Social Media. *Rev. Iberoam. Tecnol. Del Aprendiz.* **2021**, *16*, 77–83.
42. Murshed, B.A.H.; Abawajy, J.H.; Mallappa, S.; Saif, M.A.N.; Al-Ariki, H.D.E. DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. *IEEE Access* **2022**, *10*, 25857–25871. [CrossRef]
43. Pendar, N. Toward Spotting the Pedophile Telling victim from Predator in Text Chats. In Proceedings of the 1st International Conference on Semantic Computing (ICSC), Irvine, CA, USA, 17–19 September 2007; IEEE Computer Society: New York, NY, USA, 2007; pp. 235–241.
44. McGhee, I.; Bayzick, J.; Kontostathis, A.; Edwards, L.; McBride, A.; Jakubowski, E. Learning to Identify Internet Sexual Predation. *Int. J. Electron. Commer.* **2011**, *15*, 103–122. [CrossRef]
45. Nobata, C.; Tetreault, J.R.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web (WWW), Montreal, QC, Canada, 11–15 April 2016; ACM: New York, NY, USA, 2016; pp. 145–153.
46. Isaza, G.A.; Muñoz, F.; Castillo, L.F.; Buitrago, F. Classifying Cybergrooming for Child Online Protection using Hybrid Machine Learning Model. *Neurocomputing* **2022**, *484*, 250–259. [CrossRef]
47. Fadhil, I.M.; Sibaroni, Y. Topic Classification in Indonesian-language Tweets using Fast-Text Feature Expansion with Support Vector Machine (SVM). In Proceedings of the International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 6–7 July 2022; IEEE: New York, NY, USA, 2022; pp. 214–219.
48. Lestari, S.D.; Setiawan, E.B. Sentiment Analysis Based on Aspects Using FastText Feature Expansion and NBSVM Classification Method. *J. Comput. Syst. Inform. (JoSYC)* **2022**, *3*, 469–477. [CrossRef]
49. Preuß, S.; Bayha, T.; Bley, L.P.; Dehne, V.; Jordan, A.; Reimann, S.; Roberto, F.; Zahm, J.R.; Siewerts, H.; Labudde, D.; et al. Automatically Identifying Online Grooming Chats Using CNN-based Feature Extraction. In Proceedings of the 17th Conference on Natural Language Processing (KONVENS), Düsseldorf, Germany, 6–9 September 2021; pp. 137–146.
50. Ma, W.; Yu, H.; Ma, J. Study of Tibetan Text Classification based on FastText. In Proceedings of the 3rd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA), Chongqing, China, 30–31 May 2019; pp. 374–380.

51. Kocon, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydlo, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. ChatGPT: Jack of all trades, master of none. *Inf. Fusion* **2023**, *99*, 101861. [CrossRef]
52. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 4 March 2024).
53. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv* **2019**, arXiv:1911.00536.
54. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
55. Faraz, A. *Curated PJ Dataset*; IEEE Dataport: New York, NY, USA, 2023. [CrossRef]
56. Borj, P.R.; Raja, K.B.; Bours, P. On Preprocessing the Data for Improving Sexual Predator Detection: Anonymous for review. In Proceedings of the 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Zakynthos, Greece, 29–30 October 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
57. Church, K.W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162. [CrossRef]
58. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
59. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. *arXiv* **2018**, arXiv:1802.06893.
60. Dharma, E.M.; Gaol, F.L.; Warnars, H.L.H.S.; Soewito, B. The Accuracy Comparison among Word2Vec, Glove, and FastText towards Convolution Neural Network (CNN) Text Classification. *J. Theor. Appl. Inf. Technol.* **2022**, *100*, 31.
61. Nguyen, H.N.; Teerakanok, S.; Inomata, A.; Uehara, T. The Comparison of Word Embedding Techniques in RNNs for Vulnerability Detection. In Proceedings of the 7th International Conference on Information Systems Security and Privacy (ICISSP), Virtual Event, 11–13 February 2021; pp. 109–120.
62. Villatoro-Tello, E.; Juárez-González, A.; Escalante, H.J.; y Gómez, M.M.; Villasenor-Pineda, L. A Two-step Approach for Effective Detection of Misbehaving Users in Chats. In Proceedings of the CLEF (Online Working Notes/Labs/Workshop), Rome, Italy, 17–20 September 2012; Volume 1178.
63. Singla, Y. Detecting Sexually Predatory Behavior on Open-Access Online Forums. In *Research and Applications in Artificial Intelligence (RAAI)*; Advances in Intelligent Systems and Computing; Springer: Singapore, 2021; pp. 27–40.
64. Ebrahimi, M.; Suen, C.Y.; Ormandjieva, O. Detecting Predatory Conversations in Social Media by Deep Convolutional Neural Networks. *Digit. Investig.* **2016**, *18*, 33–49. [CrossRef]