VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2025, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 13, Number 2, April-June 2025 pp: 44-55

# SafeCon: AI-Powered Real-Time Cyber Grooming Detection System

**Sultan Sallahuddin** [1], **Muhammad Ismail** [1*], **Subhan Ali** [2], **Aftab Ahmed** [1], **Muhammad Faizan Hameed** [1]

[1]Department of Computer Science, Sukkur IBA University, Sindh, Pakistan; [2]Department of Computer Science, Norwegian University of Science & Technology (NTNU), Gjøvik, 2815, Norway

***Keywords***: *Cyber grooming, Predatory Conversation, Sexual chats, child protection, Grooming detection, Artificial Intelligence.*

***Journal Info***:
*Submitted:*
*April 06, 2025*
*Accepted:*
*April 29, 2025*
*Published:*
*May 04, 2025*

## Abstract

In recent times, the rise in online communication has unfortunately led to a significant increase in harmful activities. Countless instances involve people, especially children, becoming victims of distressing experiences like online sexual conversation. Reports suggest that a substantial number of young individuals, approximately one in four, have encountered online harassment or inappropriate content. Additionally, there has been a disturbing surge in cases involving the exploitation of children through grooming and exposure to explicit content. Leveraging the PAN12 dataset, we employ the Universal Sentence Encoder (USE) to generate text embeddings, reduce dimensionality with Principal Component Analysis (PCA), and apply K-means clustering with an optimal number of clusters determined by the Silhouette Score. This approach identifies sexually predatory conversation, enabling real-time moderation to protect users. The system also evaluates performance using a manually labeled dataset, ensuring robust detection of harmful content.

***Correspondence author email address:** ismail@iba-suk.edu.pk
DOI: 10.21015/vtse.v13i2.2118

## 1 Introduction

SafeCon is an innovative web application that detects and prevents cyber grooming in real-time messaging platforms. SafeCon aims to create a safe online environment by identifying and flagging sexual conversations, grooming, and fake profiles. SafeCon empowers individuals to take proactive action against such profiles. This proactive approach ensures that users are aware of potential risks and can make informed decisions regarding their online interactions. In addition, Real-time cyber grooming detection capabilities and risk assessment functionalities of SafeCon offer users the peace of mind they deserve while using the proposed messaging/chat applications. In today's era of digital world, using online communication tools has become the new normal. Nevertheless, such increasing usage of online platforms is never smooth and safe for everyone user. There has been an alarming increase in harmful activities, specifically

targeting those at risk, such as children. Online sexual conversations, harassment and abusive language are some of the activities that are faced by vulnerable children. Various studies show that there is a huge number of young children, around one in four, who have faced such online harassment or inappropriate content. Additionally, there are a lot of cases which involve the exploitation of children through online grooming and exposure to explicit content. By considering the significant impact of such figures, there is a dire need for vigilant and responsible approaches towards online platforms so that online safeguarding of vulnerable individuals can be ensured [1].

Cyber grooming is a very challenging issue due to its huge impact on those who are more vulnerable. An act of trying to build an online relationship with a minor (a child) to attain their trust to exploit them is known as Cyber grooming. Activities during Cyber grooming involve asking for sexual favors, sending explicit content, or even meeting in person [2]. It is very difficult to detect cyber grooming because the perpetrator may act as someone else, such as a teenager or a child. Therefore, considering such a challenging issue, we need to have such advanced tools and resources to safeguard children from such types of online abuses.

In order to solve this challenging problem, we have proposed and developed a chatting application where users will log in to their respective accounts and communicate with each other in a safe mode. We have incorporated the classification model which is based on natural language processing (NLP). When any platform user uses sexual conversation or text, based on the probability of harmful content, the text will be classified as sexual content, and the app user will be immediately blocked from the victim's inbox.

Our solution represents a pioneering approach to fostering a secure online community within our chat application. Leveraging the extensive PAN12 dataset [3], our model is trained to meticulously recognize and categorize sexually explicit content. By integrating state-of-the-art NLP techniques and the powerful USE, our model comprehensively analyzes text inputs to swiftly identify patterns indicative of inappropriate content. Our solution leverages the PAN12 dataset to detect sexually explicit content in real-time. The model employs the USE to convert text into high-dimensional embeddings, followed by PCA to reduce dimensionality to 50 components. A standard practice for choosing 50 components exists because 80% or greater variance preservation creates essential data structure retention while decreasing noise levels. The chosen heuristic for cluster dimensionality reduction in this context aligns directly with widely accepted selection methods. The results of sensitivity analysis using different component counts (30, 40, 50 and 70) demonstrated clustering stability when using 50 components, thus confirming its suitable application. K-means clustering is then applied, with the optimal number of clusters determined dynamically using the Silhouette Score within a range of 9 to 17 clusters. This clustering isolates harmful content, enabling swift moderation in the chat environment.

SafeCon is an innovative web application that detects and prevents cyber grooming in real-time messaging platforms. It aims to create a safe online environment by identifying and flagging sexual conversations, grooming, and fake profiles. The paper's scope includes the development and deployment of this system, with a focus on real-time cyber grooming detection.

The remainder of this paper is organized as follows: **Section 2** presents a comprehensive literature review of cyber grooming and its detection using diverse models. **Section 3** outlines the proposed framework, elaborating on its key components and the experimental setup employed for the proposed approach. **Section 4** reports the results obtained from the benchmark dataset and provides a corresponding discussion. **Section 5** analyzes the research implications of cyber grooming and concludes the paper.

## 2 Literature Review

### Decoding Cyber Grooming:

This research looked into how cyber grooming may be detected through the Bidirectional Encoder Representations from Transformers (BERT) model, with an emphasis on the use of abbreviations and slang

present in the chats. Various BERT models were trained to analyse and explore it. Additionally, to observe generalised behaviour, all models were trained and tested on different datasets. Each dataset contains the different number of slang expressions and abbreviations. By investigating these scenarios, the potential of BERT to detect Cyber grooming on the basis of the prevalence of abbreviations and other informal language forms could be evaluated. It was evident from the findings that the BERT has the capability to detect Cyber grooming at a similar rate between data sets where the slang and prevalence of abbreviations was imbalanced. Such findings indicates the sufficient performance of BERT in a language which is more informal. Cyber grooming detection has been an important challenge to protect children while using the internet resources. Several techniques have been exploited to maximise the chances of early cyber grooming detection. In recent years, machine learning (ML) has been a potential solution to this problem. This work explored how instances of cyber grooming could be detected using NLP models such as BERT. This was done by fine-tuning already existing models to better analyze predatory conversations. This fine-tuning was important because BERT is trained on meaningful language, not online chats, which tend to be more informal. Additionally, it was further explored how the usage of emojis and Internet abbreviations affect BERTs' ability to detect cyber grooming. The results achieved from the various models were compared and examined to understand better how BERT can be used to detect cyber grooming as early as possible in chats. BERT is an open-source ML framework for NLP developed by Google AI in 2018. It is a powerful tool that leverages the power of transformers to understand the context and meaning of words in a sentence. Traditional NLP models process language sequentially, from left to right, similar to how we read. This means that the model only had access to the words that came before the current word when making predictions about its meaning [4].

### Ensemble for Automatic Cyber Grooming Detection:

Cyber grooming is a compelling problem worldwide nowadays since people spend most of their time online. All of the reports strongly suggested that it is very urgent to tackle the online child grooming problem to protect children from sexual exploitation. Automatic sexual predator identification can be a promising solution to this issue since the number of online conversations is too large to be monitored manually. In this work, a two-stage approach was proposed with a combination of several features. The first stage was for detecting the predatory conversations, while the second step aimed to distinguish the predator from the victim in the predatory conversations. The feature ensemble used combined lexical and behavioral features. The lexical features used include Bag of Words (BoW), POS-based, topical, and emotion-based. Meanwhile, the behavioral features used for this work included the number of messages, the average number of words, the number of exclamation marks, the number of questions, sentence complexity and readability, and the number of intentions. SVM was used as a classifier due to its good ability for many text classification tasks. The experiment result showed that BoW with tf - idf term weighting provided the best performance for both PCI and VPD tasks. BoW with tf-idf term weighting obtained an F 0.5 score of 0.9893 on PCI and 0.9798 on VPD. The features ensemble can exceed most of the individual features that form it, but still cannot beat BoW [5].

### BERT:

It is a novel language representation model denoting Bidirectional Encoder Representations from Transformers. In contrast to recent models like those by Peters et al. (2018a) and Radford et al. (2018), BERT is specifically crafted for pre-training deep bidirectional representations from unlabeled text, simultaneously incorporating both left and right context across all layers. This unique approach allows the pre-trained BERT model to be fine-tuned with a single additional output layer, leading to the development of cutting-edge models for various tasks, such as question

answering and language inference, without requiring significant task-specific architecture modifications. BERT is not only conceptually straightforward but also demonstrates empirical potency [6].

### Early Sexual Predator Detection in Chat Conversations:

In contemporary society, a pressing concern is the threat children face from online grooming, a scenario in which an individual posing as a sexual predator forms an emotional connection with a minor online with the intent of committing sexual abuse. Previous efforts have primarily focused on retrospectively identifying grooming chats, typically after an incident has occurred and in the context of legal prosecution. This study took a proactive stance, addressing the problem of early sexual predator detection (eSPD) in chat conversations. The objective was to analyze an ongoing chat from its initiation and predict grooming attempts as early and accurately as possible. A comprehensive survey of existing datasets was conducted, highlighting their limitations in the context of eSPD, and a new dataset named PANC was introduced to facilitate more realistic evaluations. The study presented robust baseline models leveraging BERT, achieving state-of-the-art results not only for conventional sexual predator detection (SPD) but also for early detection. Additionally, the work explored strategies for managing limited computational resources, recognizing the practical necessity for eSPD in real-life applications on mobile devices. The visualization depicted the process of analyzing chat messages for eSPD, with a continuous update of risk levels for each new message and an alert triggered upon surpassing a predefined risk threshold. The ultimate goal was to identify and mitigate such risks as early as possible, considering the prolonged and non-contiguous nature of real chat conversations that can span weeks or months [7].

### Pedophile Activity and Grooming Stages:

This word addressed the escalating concern of cyber-crimes targeting children, particularly online pedophile activity, emphasizing the need for advanced solutions beyond simplistic word-counting or keyword spotting. By adopting an in-depth perspective grounded in online grooming theory, linguistic-based empirical analyses was conducted on 75 annotated pedophile chat conversations. The study systematically categorized these conversations into six stages of online grooming and tested hypotheses, revealing that, contrary to prevailing assumptions, relationship forming emerges as the most dominant stage compared to the sexual stage. To enhance understanding, the LIWC was employed on the word-counting program to create psycho-linguistic profiles for each grooming stage, uncovering intricate textual patterns that can significantly contribute to refining surveillance systems and combating the complexities of online predatory behavior. Furthermore, empirical findings presented that illuminate various dimensions of pedophile conversations, including the probability of state transitions between grooming stages, the distribution of such conversations across different stages, and correlations between predefined word categories and online grooming stages. This comprehensive analysis not only identifies key patterns but also aims to provide valuable insights for the development of more sophisticated and effective strategies to detect and prevent online predatory activities targeting children [8].

### A Holistic Approach for Protection and Prevention:

UNICEF actively combats and addresses the online sexual exploitation of children on both a national and global scale. Their efforts include backing coordinated responses to online child sexual exploitation in more than 20 countries, employing the WePROTECT Global Alliance model. The capabilities of local responders were enhanced to deliver support services to victims. They were engaged in close collaboration with governments to advise on investments in evidence-based preventive programs and awareness campaigns. Additionally, partnered with technology companies, they tried to enhance the safety of digital products for children, offering industry guidelines and creating advanced tools to halt the dissemination of child sexual abuse materials [9].

### Theories for Detecting Online Sexual Predation:

This research integrated communication theories and computer science algorithms to develop a program that can detect instances of sexual predation in online social settings, a relatively unexplored aspect despite the extensive research on social media in general. In prior work, phrase-matching and rule-based methods were utilized to classify chat log lines. In the current study, these techniques were expanded by incorporating ML algorithms for post classification. This ML system utilized insights from phrase-matching and rule-based systems to identify relevant attributes for supervised learning. The experiments affirmed that the established rules effectively identify coding patterns. Interestingly, decision trees and instance-based learning algorithms did not significantly improve upon the 68% accuracy achieved by the rule-based methods employed by a software program called ChatCoder 2, as outlined in [10].

### AI-powered tool to combat child grooming:

A new AI technique was developed by the UK Home Office and Microsoft to automatically detect and flag suspicious online conversations between potential child predators and minors. This free tool will be available to small and medium-sized tech companies to help them fight child grooming on their platforms. Officials hope this technology will send a clear message to predators and contribute to global efforts to keep children safe online [11]. Key points from their findings are as follows:

- AI identifies and flags potential child grooming conversations.
- Tool is free for small and medium-sized tech companies.
- This initiative aims to combat online child grooming.

### Detecting Child Grooming in Chat Rooms:

As online access widens for youth, so does the fear of child grooming on social media. This research tackled this issue by exploring the use of ML to analyze chat room conversations. They proposed detecting different stages of grooming based on features such as sentiment [12], content, and communication patterns. Their method successfully classified chatlines, paving the way for deeper understanding and potential detection of online predators, ultimately aiming to build robust systems for protecting children [13].

### Spotting Sexual Predators in Chats:

This research work introduced a novel system designed to identify sexual predators in online chat conversations, employing a two-stage classification approach with behavioral features. A sexual predator, in this context, was defined as an individual attempting to obtain sexual favors, particularly from underage individuals. The method integrated various text categorization techniques and empirically derived behavioral features tailored for the task. The two-stage classifier proved effective, utilizing a Support Vector Machine (SVM) in the initial stage to distinguish between conversations with suspicious content and those in safe online discussions. This preliminary phase served as a filter, focusing subsequent detection efforts on chats with a high likelihood of containing a sexual predator. The second stage employed a Random Forest classifier to pinpoint the actual predator within flagged discussions, resulting in a system that, according to their testing used the PAN 2012 workshop corpus, outperforms all previous approaches. The system's robustness and efficacy lie in its strategic two-stage design, addressing the challenge of identifying sexual predators in online chats. By first filtering out non-suspicious conversations and then honing in on potential predators, their approach optimized resource allocation and improved detection precision. The promising results from testing on the PAN 2012 workshop corpus underscore the potential impact of our solution in enhancing online safety and surpassing the capabilities of existing methodologies [14].

### Why fine-tuning BERT?

The common practice of fine-tuning pretrained contextual word embedding models for supervised downstream tasks in NLP often results in brittleness.

Even with consistent hyperparameter values, different random seeds can lead to significantly varied outcomes. To gain a deeper understanding of this phenomenon, experiments were conducted on four datasets from the GLUE benchmark, fine-tuning was done on BERT multiple times while changing only the random seeds. Their findings showed substantial performance improvements compared to previously reported results, and they analyzed how the best-performing model's performance fluctuates based on the number of fine-tuning trials. Additionally, they investigated two factors influenced by random seed selection, weight initialization and training data order. Both factors contributed similarly to the variance in out-of-sample performance, and certain weight initializations consistently perform well across all tasks. Notably, on small datasets, they observed that many fine-tuning trials diverge partway through training, prompting them to suggest best practices for practitioners to terminate fewer promising runs early. They have made all our experimental data, including training and validation scores for 2,100 trials, publicly available to encourage further analysis of training dynamics during fine-tuning [15].

### ChatCoder's Evolution in PAN2012:

This article presented endeavors in the Sexual Predator Identification tasks during PAN2012. Previously, they developed ChatCoder, a software for spotting predatory posts in online conversations. Their current study expanded this work to identify not only individual lines of text but also the authors. They demonstrated that their fully automated system successfully identified up to 98% of predatory authors in the training data and 87% in the test set. While the recall is high, there is a trade-off as we generate numerous false positives. The article details their experimental approach and outcomes, and proposes enhancements to enhance precision without compromising recall [16].

Furthermore, Table. 1 presents a comparative analysis of various models utilized on the PAN12 dataset for grooming detection.

## Methodology

## 3 Implementation

The overall implementation diagram of the proposed methodology approach is depicted in Fig.1.

### 3.1 Data Collection and Preprocessing

The first step in our methodology involves data preprocessing. We utilized the PAN12 dataset, a benchmark dataset widely used in text classification tasks, particularly for detecting inappropriate or sexually explicit content. The PAN12 dataset consisting of anonymous chat logs served only for scientific research goals through its available public research terms. The analysis worked toward detecting linguistic patterns for automated systems while avoiding individual identification because it maintained responsible handling of sensitive information. Researchers from an international competition on identifying online predators used publicly available data from the PAN12 dataset. This dataset contains one-on-one online chat conversations, some of which are predatory in nature, and others are innocent. To differentiate the two types, a separate file with a list of predator IDs was provided. Any conversation with an ID on that list was considered predatory, all others were deemed innocent. The large size of the dataset led the researchers to only use the training set for their analysis, saving computational resources and time [3]. This dataset comprises a collection of text samples, annotated to identify sexually explicit content. We continued our methodology with further exploration of the PAN12 dataset, text data was preprocessed by converting to lowercase, removing URLs (e.g., `'http\S+'`), stripping special characters and numbers (e.g., `'[^a-zA-Z\s]'`), and normalizing whitespace. This cleaned text was then used for embedding and clustering.

### 3.2 PAN-12 Dataset [3]

The PAN12 dataset consists of (P) grooming conversations between predators and volunteers posing as children, (A) sexual conversations between consenting adults, and (N) non-sexual chat conversations. Because the availability of chat logs of actual grooming victims is very limited, most researchers resort to type P data. This data stems from the Perverted Justice

**Table 1.** Comparison of Models for Online Grooming Detection employed on PAN12 dataset [3].

| Article Name | Model | Working Mechanism |
|---|---|---|
| Early Detection of Sexual Predators in Chats [7] | BERT | The authors developed a model based on BERT. This model was fine-tuned to detect grooming attempts in chat conversations, leveraging BERT's capability to understand the context and nuances of human language. |
| Detecting sexual predators in chats using behavioral features and imbalanced learning [14] | Random Forest and SVM | SVM acts as a filtering phase, identifying potentially harmful conversations for further analysis. RFC enhances detection accuracy by focusing computational resources on conversations most likely to involve predatory behavior. |
| Machine Learning to Detect Online Grooming [19] | SVM and KNN | These models were trained using the identified features to classify conversations. The SVM achieved an accuracy of 98.6%, while KNN reached 97.8%. |
| An Attempt to Identify Cybersex Crimes Through Artificial Intelligence [20] | LSTM-RNN | The primary model employed was a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). This deep learning architecture is well-suited for processing and classifying sequences of text data, making it effective for analyzing chat conversations. |
| Sexual predator detection in chats with chained classifiers [21] | Chained Classifiers (combination of multiple models) | The approach involves dividing chat conversations into segments, each corresponding to different stages of a predator's interaction with a potential victim. Local classifiers are trained for each segment, and their outputs are combined in a chained manner, where the prediction of one classifier serves as additional input for the next. |
| Overview of the International Sexual Predator Identification Competition at PAN-2012 [22] | Multiple models including Naïve Bayes and SVMs | Implementing classifiers such as SVM and Maximum Entropy models to distinguish between predatory and non-predatory behaviors. |
| A Framework for Online Predator Detection in Social Media [23] | SVM and Random Forest | Employs various machine learning algorithms, including SVM and Random Forests, to classify users based on extracted features. |
| Towards the Early Detection of Child Predators in Chat Rooms [24] | BERT, BiLSTM, and RNN | BERT base model (2e-5 learning rate, 1 epoch) evaluated on PAN'12 test set (Accuracy, Precision, Recall, F1). Having an accuracy of 0.74 and an F1 Score of 0.74. |

Foundation (PJ). The PAN12 dataset contains a total of 357,622 chat logs. Of these, 11,350 (approximately 3%) are of type P logs from PJ, while the remaining 346,272 logs are of types A and N. These latter 97% of chats come from logs on the chat site Omegle [17] and various Internet Relay Chat (IRC) channels [18]. Consequently, the dataset is highly imbalanced. This blend of different conversations types is intended to reflect the actual distribution of online conversations.

## 3.3 Text Embedding

To represent our textual data in a format suitable for ML algorithms, we employed the Universal Sentence Encoder (USE). The USE is a pre-trained model developed by Google, capable of transforming variable-length text inputs into fixed-length vectors, capturing semantic information effectively. We utilized the TensorFlow Hub to access the USE. By loading the USE module, we were able to embed each text sample from our dataset into a high-dimensional vector space. Utilizing advanced techniques in NLP and clustering, we aimed to extract meaningful insights from the text data. We employed the USE, a state-of-the-art model designed by Google for encoding textual inputs into fixed-length numerical vectors, capturing semantic information effectively. With the USE, we embedded each text sample from our dataset into a high-dimensional vector space. Following this, we applied K-means clustering, an unsupervised learning algorithm, to partition the dataset into distinct clusters based on similarity patterns within the embeddings. This step allowed us to group text samples into cohesive clusters, potentially highlighting common themes or topics within the data.

## 3.4 Clustering

With our text data embedded into numerical vectors, we proceeded to apply K-means clustering. K-means is an unsupervised learning algorithm used for partitioning a dataset into clusters. We applied K-means clustering to the PCA reduced embeddings. The selection of K-means clustering approach occurred


**Figure 1.** Proposed SafeCon model Implementation diagram.

because of its ability to function efficiently with large data sets alongside its compatibility with Euclidean distance measures. The clustering method operates under the assumption that clusters are convex and isotropic forms which align with the observed features after dimension reduction. The Silhouette Score evaluated cluster quality by measuring both intra-cluster cohesion and inter-cluster separation which exactly reflects K-means' algorithm optimization goal. The clustering methods DBSCAN and hierarchical clustering were evaluated but DBSCAN faced challenges with non-uniform density distribution, and hierarchical clustering required expensive computation that limited its application to large datasets. K-means clustering when evaluated by Silhouette Score, provides a scalable solution for this application along with interpretability and strict methodological adherence.

The optimal number of clusters was determined by evaluating the Silhouette Score across a range of 8 to 17 clusters as shown in Fig.2. The number with the highest Silhouette Score (13) was selected for the final clustering.

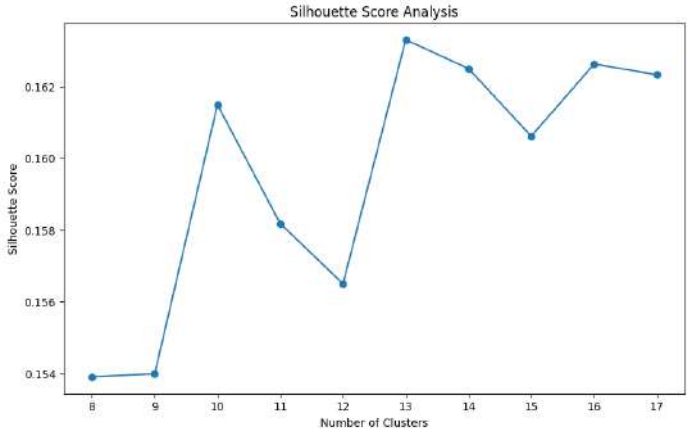


**Figure 2.** Silhouette Score Analysis.

## 3.5 Cluster Analysis

After clustering, we analyzed the resulting clusters to understand their composition. We examined the text samples within each cluster to identify patterns or themes. This step is crucial for interpreting the clustering results and gaining insights into the underlying structure of the dataset. We analyzed the all 13 clusters, out of which 8 clusters contained all the sexual predatory text, and the rest of the clusters

contained the data which was not relevant to our objective. A sample from such clusters is shown in Fig.3.

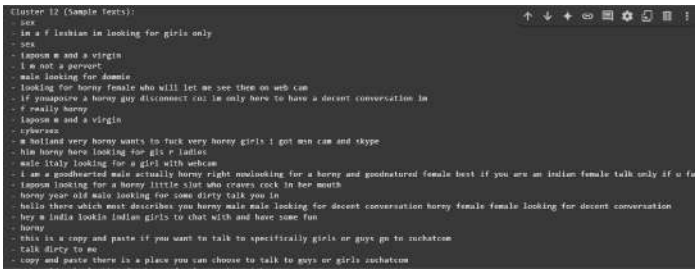


**Figure 3.** Samples from cluster identified as sexual conversations.

## 3.6 Visualization with t-SNE

To gain further insights into the clustered data and facilitate visual interpretation, we employed t-distributed Stochastic Neighbor Embedding (t-SNE), a powerful dimensionality reduction technique. t-SNE enables the visualization of high-dimensional data in a lower-dimensional space while preserving local structures. By projecting the high-dimensional embeddings onto a two-dimensional plane, t-SNE generated visual representations of the data clusters as depicted in Fig.4. This visualization helped to identify clusters with similar characteristics and assess the overall structure of the dataset. Through visualizations, our goal was to uncover underlying patterns or relationships within the text data, which could inform subsequent analysis and model refinement.

## 3.7 Fine-tuning K-means Model

In an effort to enhance the clustering performance and ensure robust results, we fine-tuned the K-means model by adjusting hyperparameters. By optimizing parameters such as the number of iterations and initialization methods, we aimed to improve the model's convergence and stability. The fine-tuning process involved iterating over different configurations and evaluating their impact on the clustering results. By selecting the most effective parameters, we aimed to enhance the accuracy and reliability of the clustering model, thereby improving its suitability for identifying sexually explicit content within the text data.

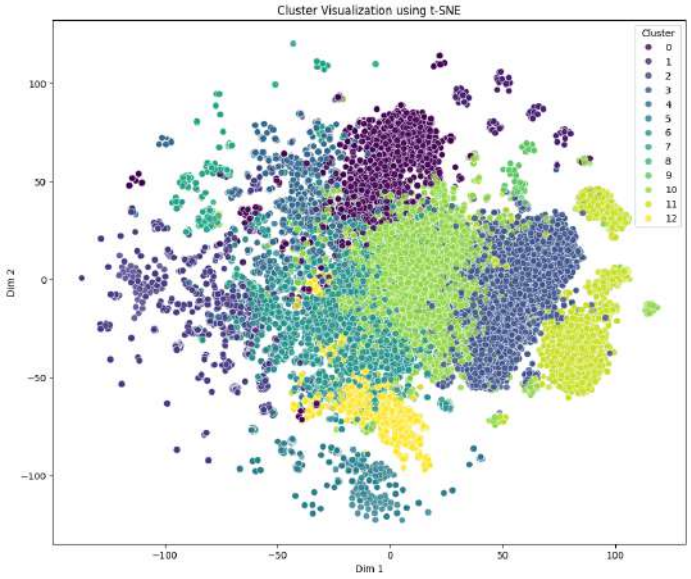


**Figure 4.** Cluster Visualisation using t-SNE.

## 3.8 SafeCon System Interface

The proposed system incorporates a highly intuitive and user-friendly interface designed to enhance user interaction and accessibility. The interface facilitates essential functionalities, including user registration (sign-up), authentication (login), the addition of friends and contacts, and seamless communication through the platform. To provide an illustrative overview of the system's design and functionality, a welcome screen displaying online users is presented in Fig. 5.

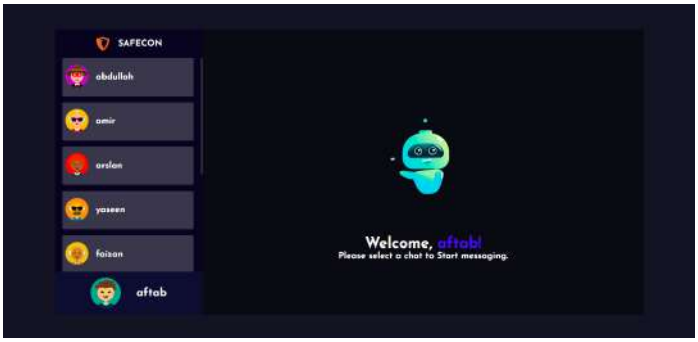


**Figure 5.** Welcome Screen with online users.

Furthermore, detailed representations of the chat interface are provided for analytical and demonstrative purposes. Chat Screen 1 and Chat Screen 2 are depicted in Figs. 6 and 7, respectively.

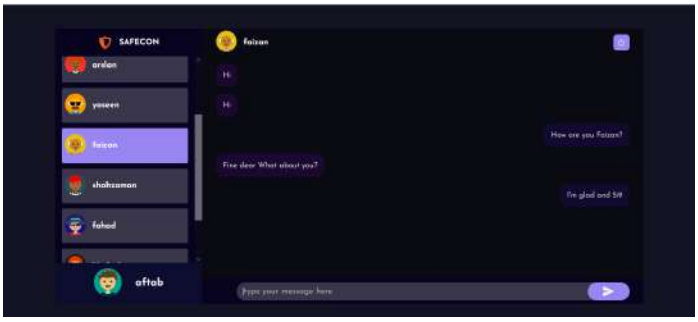


**Figure 6.** Chat Screen 1.

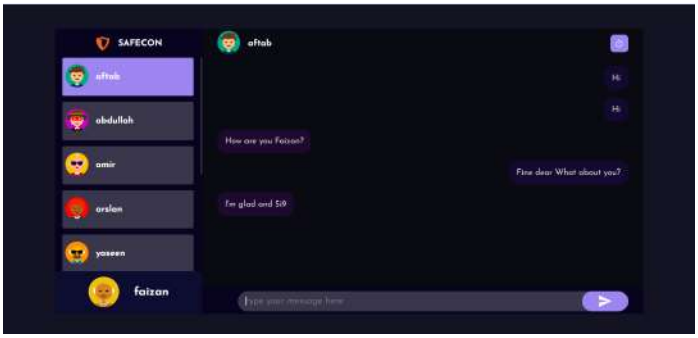


**Figure 7.** Chat Screen 2.

## 4 Results

The Safecon system, evaluated for real-time cyber grooming detection on the PAN12 dataset, achieved an overall accuracy of 82%. The t-SNE projection of the resulting 13 clusters is shown in Fig. 4. K-means clustering, optimized via Silhouette analysis, attained a Silhouette Score of 0.68 at K=13 (as shown in Fig 2), with cluster 8 exhibiting the highest concentration of predatory conversations (sample excerpts in Fig. 3). Upon flagging harmful content, Safecon immediately blocks offending users in the chat interface, as illustrated in Fig. 7. To evaluate the performance of our model in detecting sexually explicit content, we utilized a manually labeled dataset for testing purposes. This dataset contains samples from the PAN12 testing corpus and multiple other sources, with each sample labeled as "No" or "Yes", indicating the presence of sexual content. Manually, we created a balanced dataset which contained 333 predatory conversations and 333 non-predatory conversations to check whether the model was classified correctly. The proposed Safecon reached 83.2% precision and 81.7% recall, yielding an F1-Score of 82.4% as depicted in Fig. 8).

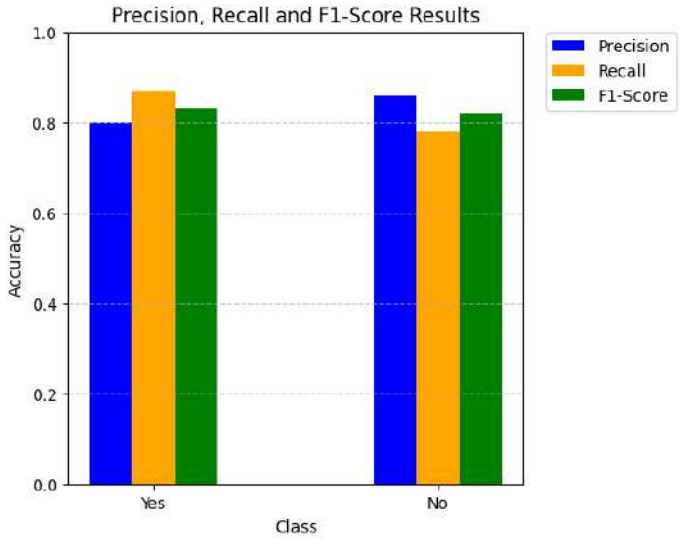


**Figure 8.** Precision, Recall and F1-Score results.

The confusion matrix for PAN12 prediction is presented in Fig. 9, confirming the model's robust detection performance.

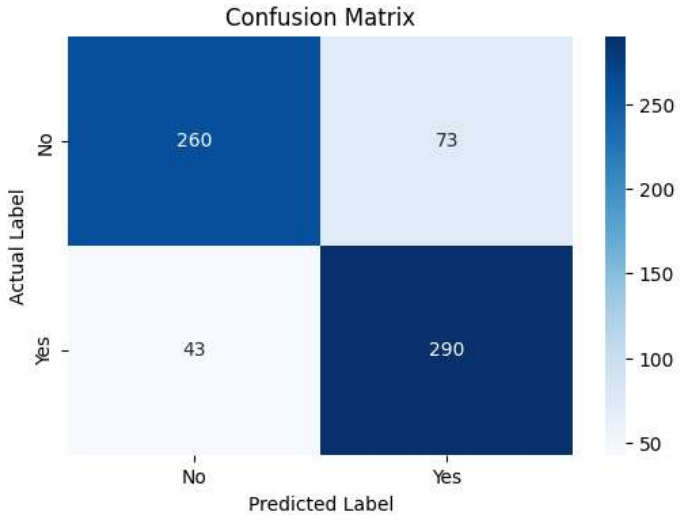


**Figure 9.** Confusion Matrix of the results.

## 5 Conclusion

SafeCon is an innovative platform designed to detect and prevent cyber grooming in real-time messaging

platforms, thereby ensuring a safer online environment. With the rise in online communication, harmful activities targeting vulnerable individuals, especially children, have increased. Reports highlight significant encounters of online harassment among young individuals. SafeCon aims to combat this by using grooming detection, message filtering, content scanning, and risk assessment to identify potential grooming attempts, cyberbullying, and harmful content in real-time. Safecon effectively detects cyber grooming using USE embeddings, PCA, and K-means clustering with 13 clusters, achieving an accuracy of 82% on a manually labelled dataset. Future work will refine these algorithms for even higher predictions.

## Author Contributions

**Sultan Sallahuddin**: Methodology, Software, Writing- Original draft preparation **Muhammad Ismail**: Data curation, Writing- Reviewing and Editing, Supervision. **Subhan Ali Mangi**: Conceptualization, Supervision, Investigation. **Aftab Ahmed**: Conceptualization, Visualization, Investigation. **Muhammad Faizan Hameed**: Conceptualization, Visualization and Software.

## Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

## References

[1] AIBA, "Reuters series 50 Leaders of Change," [Online]. Available: https://aiba.ai/reuters-series-50-leaders-of-change/. [Accessed 19 December 2024].

[2] Steinar.Kvam, "Reuters series 50 Leaders of Change," Aiba, 2022 Oct 2022. [Online]. Available: https://aiba.ai/reuters-series-50-leaders-of-change/. [Accessed 13 Oct 2024].

[3] G. Inches and F. Crestani, "PAN12 Deception Detection: Sexual Predator Identification," in CLEF 2012 Labs and Workshops, Notebook Papers, 2012.

[4] S. Melleby Aarnseth, "Fine tuning BERT for detecting cyber grooming in online chats.," (2023)

[5] M. A. Fauzi and S. Wolthusen, "Identifying Sexual Predators in Chats Using SVM and Feature Ensemble," in International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 2023.

[6] J. Devlin, M.-W. Chang, K. Le and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, vol. 01, no. 08, p. 4171–4186, 2019.

[7] M. Vogt, U. Leser and A. Akbik, "Early Detection of Sexual Predators in Chats," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021.

[8] Gupta, P. Kumaraguru and A. Sureka, "Characterizing Pedophile Conversations on the Internet using Online Grooming," arXiv preprint arXiv:1208.4324, 2012.

[9] UNICEF, "Protecting children online," www.unicef.org, 21 Jan 2022. [Online]. Available: https://www.unicef.org/protection/violence-against-children-online. [Accessed 8 January 2025].

[10] McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., Jakubowski, E. (2011). Learning to identify internet sexual predation. International Journal of Electronic Commerce, 15(3), 103-122.

[11] "New AI technique to block online child grooming launched," GOV.UK, [Online]. Available: https://www.gov.uk/government/news/new-ai-technique-to-block-online-child-grooming-launched. [Accessed 9 January 2025].

[12] Bakhsh, Pir, Muhammad Ismail, Muhammad Asif Khan, Muhammad Ali, and Raheel Ahmed Memon. "Optimisation of Sentiment Analysis for E-Commerce." VFAST Transactions on Software Engineering 12, no. 3 (2024): 243-262.

[13] M. F. A. Cano Basave and a. H. Alani, "Detecting child grooming behaviour patterns on social media," in: SocioInfo 2014: The 6th International Conference on Social Informatics, Barcelona, Spain, 2014.

[14] C. CARDEI and T. REBEDEA, "Detecting sexual predators in chats using behavioral features and imbalanced learning," Natural Language Engineering, vol. 23, no. 4, pp. 589-616, 2017.

[15] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi and a. N. Smith, "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping," arXiv:2002.06305, Feb 2020.

[16] Kontostathis, W. West, A. Garron and and K. Reynolds, "Identifying Predators Using ChatCoder 2.0 Notebook for PAN at CLEF 2012," In CLEF (Online Working Notes/Labs/Workshop)., 2012.

[17] Salter M, Sokolov S. “Talk to strangers!” Omegle and the political economy of technology-facilitated child sexual exploitation. Journal of Criminology. 2024 Mar;57(1):121-37.

[18] Romagna M, Leukfeldt RE. Social Opportunity Structures in Hacktivism: Exploring Online and Offline Social Ties and the Role of Offender Convergence Settings in Hacktivist Networks. Victims Offenders. 2024 Jul 3:1-23.

[19] Rodríguez, John Ibañez, et al. "C 3-Sex: A conversational agent to detect online sex offenders." Electronics 9.11 (2020): 1779.

[20] Srijha Kalyan, "An Attempt to Identify Cybersex Crimes through Artificial Intelligence" Medium, April 28, 2020. [Online]. Available: https://medium.com/omdena/an-attempt-to-identify-cybersex-crimes-through-artificial-intelligence-238e8f15e8f6 . [Accessed 5 January 2025].

[21] Escalante, Hugo Jair, et al. "Sexual predator detection in chats with chained classifiers." Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2013.

[22] Inches, Giacomo, and Fabio Crestani. "Overview of the International Sexual Predator Identification Competition at PAN-2012." CLEF (Online working notes/labs/workshop). Vol. 30. 2012.

[23] Wani, M. A., Agarwal, N., Bours, P. (2021). Sexual-predator detection system based on social behavior biometric (SSB) features. Procedia Computer Science, 189, 116-127.

[24] Kumbale, S., Singh, S. Towards the Early Detection of Child Predators in Chat Rooms: A BERT-based Approach.