# Evaluating a Non-platform-specific OCR/NLP system to detect Online Grooming

**Jake Street, Dr Funminiyi Olajide**
Nottingham Trent University, Nottingham, United Kingdom

jake.street02@ntu.ac.uk
funminiyi.olajide@ntu.ac.uk

**Abstract:** Online Grooming is a social engineering attack in which the attacking party uses deceptive practices for sexual gratification. The targets of these attacks can vary in demographics however in most cases the target is children, with most of these attacks occurring on social media platforms. As well as the illegality of these attacks in the UK and US, children who experience these attacks are at a higher risk of self-harm or having suicidal thoughts. Due to the deployment of new social media platforms/features any implementation that is made specific to a certain feature/platform is likely to be outdated/ineffective upon release, due to the volatility of the methods/tactics used. Therefore a non-platform specific implementation has been considered within this investigation. From a preliminary analysis, it was concluded that there was an average true positive detection rate of 71% from using optical recognition and natural language processing across three different social media platforms. It is suggested that implementing this text extraction and processing method alongside a 'category-based' machine learning algorithm, a solution that can identify online grooming can be developed that considers the 'real world complexities' of this attack.

**Keywords:** Natural Language Processing, Machine Learning, Optical Character Recognition, Online Grooming

## 1. Introduction

Individuals that carry out social engineering attacks have a wide range of motives as to why they are carrying out their attack. Some social engineers have similar motives to attackers that would take a more technical-based approach, with these attacks often being targeted at commercial organisations. The motives for conducting these attacks are likely to be for financial gain, to disrupt and cause harm to the organisation (possibly a disgruntled employee), or a competitor of the organisation aiming to gain a competitive advantage.

These social engineering attacks targeted at organisations follow the hypothesised 'CIA triad' of information security (Samonas, 2014) in which the 'CIA triad' refers to Confidentiality, the principle in which only authorised entities can access a given piece of information; Integrity, the principle in which information is not modified without proper authorisation; and Availability, the principle in which information is available upon demand by an authorised entity.

Table 1 demonstrates the different social engineering attacks targeted at organisations that specifically relate to each of the principles of the 'CIA triad'.

**Table 1: Social Engineering Attack Methods that Affect each Element of the 'CIA triad'**

| 'CIA triad' Principle | Social Engineering Attack Method |
|---|---|
| Confidentiality | Social biases/asking questions |
| Integrity | Requests while impersonating an authority figure |
| Availability | Theft of infrastructure (to affect a service) |

However with social engineering attacks unlike most technical attacks the target that is selected determines the likelihood that the attack will be successful, the likelihood that the attack will be identified by the victim, and the severity (risk) to the attacker if the attack were to be identified by the victim.

Because of these factors victim selection is one of the main considerations for an attacker when conducting one of these attacks, with attackers looking for specific personal attributes when selecting their target or evaluating the likelihood of success/risk of the attack.

Alike a vulnerable computer system, individuals that have a high attack success rate based upon their attributes should be referred to as 'Vulnerable individuals' with each of these vulnerabilities often having their own

corresponding exploit/attack. Some potential social engineering exploits/attacks for each of these personal vulnerabilities are shown in Table 2.

**Table 2: Personal Attributes with a Perceived Increased Attack Success Rate**

| Vulnerability/Personal Attribute | Social Engineering Attack |
|---|---|
| Naivety | All Attacks |
| Fear/Anxiety | Tech support scams |
| Technical Awareness | Tech support scams, Phishing emails |
| Lack of Awareness of Sexual Exploitation | Online Grooming, Relationship Scams |

### 1.1 Context

This investigation will be observing the Online Grooming social engineering attack which is defined as the process in which an adult forms a sexually abusive relationship with a child using technology (Lorenzo-dus, 2017). These attacks have been found to be increasing in frequency over time (Bentley, 2018) despite an increased focus on education for children surrounding this (Department for Education, 2019) in the UK. Therefore due to this perceived lack of effectiveness of education to mitigate these attacks, a greater focus on technical solutions should be considered.

## 2. Literature Review

Online Grooming from a cybersecurity approach is in its infancy within Literature, due to Online Grooming being predominantly seen within scope of the domain of Psychology and Social Sciences as opposed to Cybersecurity and Social Engineering. This should be considered when analysing this literature so that the findings presented can be best considered for a cybersecurity approach.

### 2.1 Psychological Considerations

There are many different methods/strategies that are used by online groomers when conducting their attack which makes creating an effective solution technically complex. Despite this variety of approaches taken by attackers, there remains constants/'phases' between all OG attacks that perpetrators tend to follow throughout communication with the target. These phases are based on a chatroom context so therefore it should be noted that these phases are likely not generalizable to other contexts/mediums that are used within OG attacks (e.g. within a social media live streaming service).

Initiation Phase: The purpose of this phase is to determine the suitability of the target. A suitable target will depend on the attacker's preference in terms of age and gender. This is often the first communication between the attacker and the target which is done to scope the target for personal details.

Determining Method Success. Once the target has been deemed suitable, the groomer may or may not attempt to ask questions to ascertain what method would be best to execute the attack. This commonly involves having 'friendship-forming' conversations, which is often used to identify what the target is likely to respond positively to (O'Connell, 2003).

Determining Attack Risk Phase: This is based upon the groomer's risk tolerance they may or may not 'complete a risk assessment' to determine the risk that they are likely to face from executing the attack (O'Connell, 2003). This risk assessment shall be used to determine the probability of getting detected as well as the severity of detection.

Sexual Gratification Phase: The final stage of the grooming process is executing the attack, often with the intent of sexual gratification for the attacker. This can include requests from the attacker or more manipulative strategies such as blackmail.

#### *2.1.1 Victim-centric Approach*

While many victims of Online Grooming and sexual abuse feel it is their fault for the attack occurring (Patel, 2021), this is clearly not the case. However, there are a variety of attributes and behaviours that appear at a higher frequency for individuals who have experienced one of these attacks as opposed to the average person.

It should be noted that without a longitudinal study of children throughout growing up, it cannot be concluded that these attributes/behaviours have an impact on the chance of an attack occurring or if these attributes/behaviours are developed from experiencing the attack in the first place. This latter point does have some validity, as Patel (2021) observed that individuals that have experienced this form of abuse are more likely to seek similar abusive scenarios in a process known as 'revictimization'.

'Vulnerable behaviours' in this context refers to activities that a child does which will increase the probability of an attack occurring and/or activities that increase the severity of an attack. For example, a vulnerable behaviour that increases the probability of an attack occurring would be going on chatroom or other 'person discovery' websites as 10-20% of children that visit these websites receiving sexual messages from adults (Finkelhor, 2000).

An activity deemed a 'Vulnerable behaviour' that increases the severity of the attack would be a child using a livestream feature within a social media platform that allows comments from watchers. This increases the severity of a potential attack as the child is put under a time constraint to respond to comments from watchers as well as the potential for multiple attackers to gang up on one target.

#### *2.1.2 Perpetrator-centric Approach*

It is suggested that some perpetrators, which O'Connell (2003) categorised as "cyber rapists", are not concerned about mitigating their risk, by carrying out a 'risk assessment'. Instead, it is suggested that a 'hit and run' tactic is used, which involves outright aggressive acts towards the victim. This strategy aims to capitalise on the fear of the victim, as opposed to a strategy in which the victim is given more control. These attackers rely on any possible anonymity they may have and constant threats to the victim. It is assumed from the description given by O'Connell (2003) that when an attacker deems a situation as too risky, they attempt to mitigate the attack by "running" and ceasing all communication, as well as possibly using destructive measures or further threats to the victim which can be attributed to the 'Damage Limitation' phase (O'Connell, 2003).

In pretty much all cases, grooming is done purely for sexual gratification. However as deemed by Wood (2013) paedophiles do not feel "wholly adult", so it could be hypothesised that some grooming attacks could be done for social reasons.

Conte (1989) discovered that most (19 in 20) groomers had a "secondary deviance", in addition to their main grooming behaviour. Therefore, it can be likely that these may be combined in their attacks, and thus should be considered when designing a solution to detect these attacks.

### 2.2 Current Solutions

This section shall describe the solutions that have been implemented to attempt to mitigate Online Grooming attacks. Some of these implementations may be more widespread than others, with some solutions described here being hypothetical.

It should be noted that these solutions are likely to differ in appropriateness based on if they are to be implemented within a school or home setting: with a school policy being more focused on security and well defined, and with a home policy being more focused on usability and is often undefined.

#### *2.2.1 Website Blocking (Allow list & Deny list)*

This solution involves stopping a Web browser from accessing a list of websites (in terms of the 'Deny list' strategy) or only allowing the Web browser to access a list of websites (in terms of the 'Allow list' strategy). These can be implemented to stop a child from accessing social media or other platforms that paedophiles tend to operate on.

The issue with implementing this is that the child is likely to not comply with these restrictions and it is easy for the child to get round these restrictions as they work at browser level (so use of a different browser/device would circumvent this mitigation method) or by finding a communication platform that has not been denied by the list.

#### *2.2.2 Activity Monitoring*

Activity Monitoring refers to the recording information about the utilisation of software on a system and the specific activities carried out on the software. Attributes that tend to be recorded within this include duration of an activity, files downloaded, messages sent and received, and a timestamp of the activity (Netnanny.com,

2022). This information can be used to detect if an Online Grooming attack is taking place however there are some significant limitations to this. Firstly, if there is no further automation, for a parent/guardian to be able to detect an Online Grooming attack it would take a significant amount of human processing, as each activity would need to be looked through in detail, which would be infeasible for most guardians. Secondly, for this implementation to be effective at helping a parent/guardian detect an online grooming attack, this is likely to be deemed an invasion of the child's privacy and could aid an abusive parent/guardian. This is likely to also cause the child to not comply with the system and look for methods to circumvent it.

#### *2.2.3 Word Blocking*

Word Blocking refers to software that can parse on-screen text on specific websites and compare this text to a list of dangerous words/phrases that refer to a variety of topics that might not be suitable for children. Following a detection of one of these dangerous words/phrases a variety of actions can occur, but most commonly one of the following; the child will not be allowed to access the website, the words/phrases are removed from the html (client-side), or the child will be able to access the website but the parent would be informed of the threat (via a screenshot of the threat or relaying the text in question).

While this implementation can allow for the real-time detection and/or blocking of a threat, this solution suffers from the general complexities of being able to detect online grooming attacks (namely being able to establish the context of a communication). This solution also suffers in terms of being functional across multiple websites, as each platform is likely to require some fine tuning with the text that should be analysed by the software (to reduce the frequency of false positives).

#### *2.2.4 Policy Publication*

Policy publication refers to the non-technical mitigation method in which users on a given network/device/service agree to follow a specific set of rules. These policies are commonly implemented in schools (Cramer, 2010) as schools tend to have a clear definition as to what content/services are appropriate for their students to access and what is not. However, in most cases, this is not the same within the home environment due to the relaxation of restrictions as a child gets older e.g. a child being able to create a social media account when turning thirteen. It should be noted that despite these service-level policies, many children will use a service when they are not of appropriate age (Office of Communications, 2017).

While these policies do not detect or stop an attack from occurring, these can stop children engaging in vulnerable behaviours on networks/devices with these policies applied. The success of these policies are based on the mutual respect of the parties involved in the policy, as if one party feels that the policy is unfair upon write up or if a party does not comply to the rules of the policy then it is not likely to be respected by all parties.

#### *2.2.5 Education Methods*

Education methods refer to the process of educating both children (Department for Education, 2019) and their parent/guardian on vulnerable behaviours and how to identify online grooming (from both the child's and parent's perspective). While parents tend to be aware of Online Grooming as an issue that affects children, children tend to not be aware of the complexities of why adults engage in paedophilic behaviour and the risks that this poses to them.

Without providing children with detailed information about these attacks, which would raise significant ethical concerns, it is difficult to rely on this method to ensure that children will firstly identify the attack and secondly know the risks that face them because of it.

## 3. Design

### 3.1 Design Scope

Within creating a solution to detect online grooming attacks, a variety of factors need to be considered to ensure that an appropriate solution is made for a specific real-world use case. The use case in question for this investigation is for a solution that will be implemented within the home environment, with the assumption that the websites/platforms do not allow direct text-extraction from an API or the website's HTML. Additionally the child shall only access these websites through a desktop computer (as opposed to a phone or tablet devices).

It is important to remember the real-world context of an implementation of this nature, as there are many ethical challenges within this area. A significant challenge with implementing an online grooming detection system is protecting children from abuse from the parent/guardian. This is because, if a 'screenshot' feature is implemented within the solution, a parent/guardian alongside the attacker may be able to blackmail the child. Additionally, the parent may find out information about their child which could cause harm e.g. the system exposing the sexuality of the child. Another real-world consideration to make is the use of 'blocking' when a detection is made, to attempt stop an attack occurring in real time, however the issue with this is that it is likely to annoy users if there is a false positive and will therefore compromise the usability of the system. These factors must be considered within the solution, Figure 1 demonstrate the 'trade offs' that need to be considered when designing a software solution for online grooming.

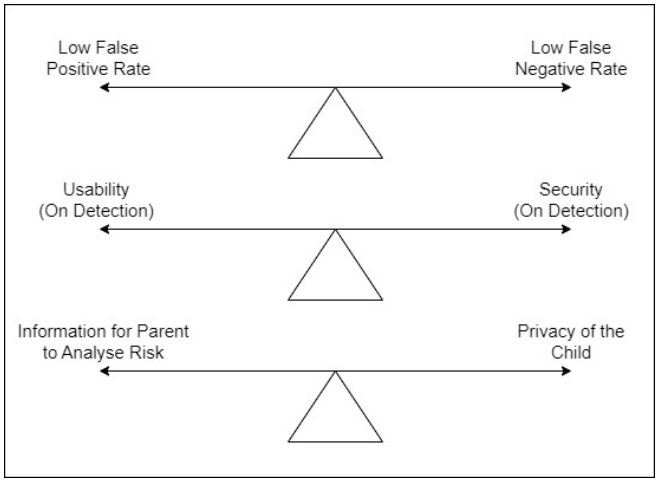


**Figure 1: Design 'trade-offs' for Online Grooming Software Solutions**

## 3.2 Design

Based upon the analysis of the current solutions it is deemed that implementing a 'Word Blocking Regular Expression' solution would be most effective considering the design scope. This will allow for some level of Online Grooming phrase detection to occur while being able to maintain a good level of privacy for the child (as opposed to using an 'Activity Monitoring' solution).

A design consideration that should be made is if the system should incorporate networking as there are some benefits in terms of features that could be implemented, these considerations are shown within Table 3.

**Table 3: Advantages between an Offline and Networked Design**

| Advantages of an offline solution | Advantages of a networked solution |
|---|---|
| Does not require an internet connection to be functional. | Parents can access information from anywhere with Internet access. |
| Does not affect Internet speed, for other services/applications. | Ability to implement notifications for the parent |
| | Ability to instantly implement an updated word/phrase blocking list for all users. |

Another complexity that needs to be considered within this use case is how to extract the text from the websites if an API or HTML parsing cannot be used. To be able to do this, screenshots of the current activity at a set regular interval in combination with an Optical Character Recognition (OCR) module shall be used. By processing the screenshots of the online activities done through OCR a string of text will be outputted.

This string of text will then be processed to ensure that any artefacts within the string, based upon misinterpreted features of the screenshot, are removed. Then following this these words will follow a splitting process to allow for further processing and for an ultimate decision on if the given image contains Online Grooming material. The individual words will go through Stemming and Lemmatization, this is to ensure that

phrases that use slightly different wording are normalised to reduce the number of false negatives (e.g. “word”, “words”, “worded”, and “wording” would all be Stemmed to “word”). Following this the individual words and phrases can be compared against the blocked words and phrases and then if a match is made a given action can be performed based upon the given use case (this can range from ‘locking’ the computer to sending a notification to the parent of a potential threat). This process is detailed in Figure 2.

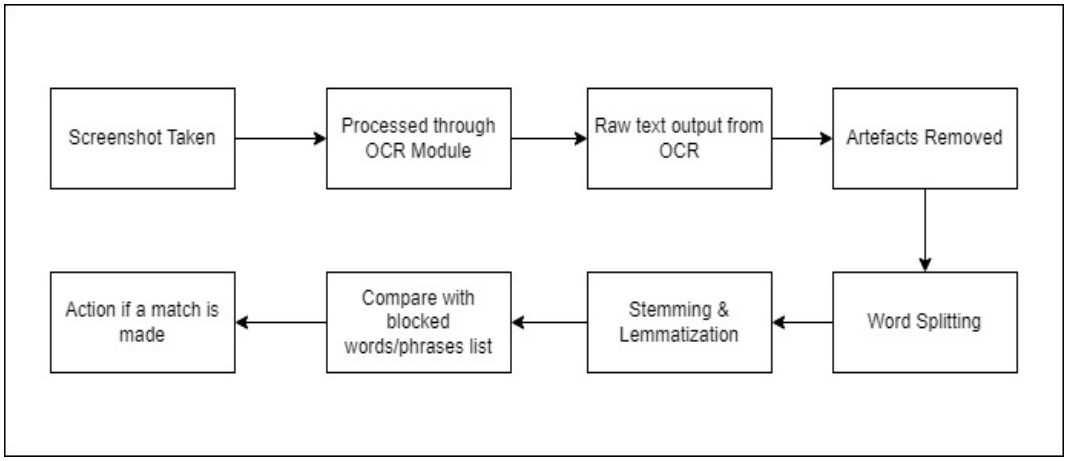


**Figure 2: High-level Description of Solution**

## 4. Methodology

The research philosophy behind this investigation is a proof of concept study to identify if using OCR is a feasible technique for text extraction for the purpose of identifying Online Grooming attacks over multiple social media platforms, or if ‘direct’ text extraction methods (such as APIs and HTML parsing) should be used.

Within this study only one OCR piece of software shall be evaluated, further research should investigate different OCR solutions to identify if there is a significant difference in text extraction for screenshots of websites.

In addition, this investigation shall only consider one screen size (23 inch). Other screen sizes should be investigate in further research to identify how this impacts accuracy, as well as observing how ‘mobile versions’ of websites affect this accuracy.

This system shall be tested using a random sample of 40 chatroom transcripts that were obtained from a ’honeypot’ Online Grooming study in the United States , sourced from ChatCoder.com (2022).

The testing of this system shall be conducted over three social media platforms: ‘Facebook’ from Meta Platforms; ‘Twitter’; and ‘e-Chat’, an anonymous unmoderated chatroom service where users can set up their own chatrooms.

Messages from ChatCoder.com (2022) will be sent from the ‘Attacker’s Computer’ to the ‘Child’s Computer’, on a variety of social media platforms, which shall be running the detection software described in Figure 2.

The Null hypothesis that shall be tested within this investigation is as follows: “There is no significance in the proportion of true positive detections to false negative detections while using OCR text extraction with NLP analysis for a specific social media platform”.

## 5. Results

Table 4 shows the true positive rate, false negative rate, and if the test null hypothesis is accepted or rejected for each of the social media platforms tested. As well as a collated, overall, result over all tests.

Within this investigation a ‘Correct Detection’ refers to an action, in this case redirecting the child’s PC to a certain website, occurring within a period of 30 seconds of a message that should be blocked by the system being sent to the child’s PC. Conversely, a ‘Missed Detection’ refers to the action occurring greater than 30 seconds from when the messages was sent or if the action did not occur from the message being sent.

**Table 4: Investigation Results**

| Test Conducted | Correct Detections (True Positive rate) | Missed Detections (False Negative rate) | Outcome |
|---|---|---|---|
| Twitter | 18 | 12 | Accept Null |
| Facebook | 20 | 10 | Accept Null |
| e-Chat | 26 | 4 | Reject Null ($p < 0.05$) |
| All | 64 | 26 | Reject Null ($p < 0.05$) |

These results show that out of the three social media platforms that was tested, only one test ("e-Chat") rejected the Null hypothesis ($p < 0.05$). The collated results across all three tests also rejected the Null hypothesis ($p < 0.05$).

## 6. Discussion

From these results it can be concluded that there is some significance in true positive detection using OCR, and while this result is somewhat promising some considerations should be made surrounding it.

Firstly, despite the statistical significance within the 'e-Chat' and 'All' tests the level of significance was low compared to an acceptable level for this use case. This is because there are clearly repeatable issues with using OCR text extraction in this context as even on a website that has a good density of text vs. images/other media, such as 'e-Chat', there were still some frequency of Missed Detections.

It can be assumed that the frequency of Missed Detections from text extracted through bespoke-to-service HTML parsing is close to zero, therefore this frequency of Missed Detection is too excessive for this use case within this implementation.

Another element that needs to be considered within the real-world implementation of an OCR/NLP solution to Online Grooming is the False Positive rate. This was not considered within this investigation, as it was out of scope, however it is likely that this will cause some usability impact on users if this were to be implemented at current.

Overall, it is suggested that further research is conducted on the effectiveness of an OCR/NLP solution on mobile devices as it is hypothesised that the results obtained would have greater ratio of true positives vs. false positives compared to the ratio obtained in this study due to less potential OCR artefacts within the text capture, reduced processing times due to smaller images, and greater OCR conversion accuracy due to an increased text size (in relation to the screen/image size).

## References

Bentley, B. C. G. G. H. L. M. O. P. P. S. S. V. W., 2018. How Safe Are Our Children: National Society for the Prevention of Cruelty to Children. [online] Available at: <https://careleaverpp.org/wp-content/uploads/2018/07/how-safe-children-2018-report.pdf> [Accessed 9 October 2022].

Chatcoder.com. 2022. ChatCoder HomePage. [online] Available at: <https://chatcoder.com/index.html> [Accessed 15 October 2022].

Conte, J.R., Wolf, S. And Smith, T., 1989. What Sexual Offenders Tell Us About Prevention Strategies. Child Abuse & Neglect, 13(2), pp.293-301.

Cramer, M. and Hayes, G., 2010. Acceptable use of technology in schools: Risks, policies, and promises. IEEE Pervasive Computing, 9(3), pp.37-44.

Department for Education, 2022. Teaching Online Safety in Schools. [online] Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/811796/Teaching_online_safety_in_school.pdf> [Accessed 9 October 2022].

Finkelhor, D., Mitchell, K.j. And Wolak, J., 2000. Online Victimization: a Report on the Nation's Youth.

Lorenzo-Dus, N. and Izura, C., 2017. " cause ur special ": Understanding trust and complimenting behaviour in online grooming discourse. Journal of Pragmatics, 112, pp.68-82.

Netnanny.com. 2022. [online] Available at: <https://www.netnanny.com/> [Accessed 15 October 2022].

O'Connell, 2003. A Typology of Child Cybersexploitation and Online Grooming Practices. [Online] Available at: http://image.guardian.co.uk/sys-files/Society/documents/2003/07/17/Groomingreport.pdf [Accessed 13 October 2022].

Office of Communications, 2017. Children and Parents: Media Use and Attitudes Report. [online] Available at: <https://www.ofcom.org.uk/__data/assets/pdf_file/0020/108182/children-parents-media-use-attitudes-2017.pdf> [Accessed 7 October 2022].

Patel, P., 2021. Tackling Child Sexual Abuse Strategy 2021. [online] Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973236/Tackling_Child_Sexual_Abuse_Strategy_2021.pdf> [Accessed 9 October 2022].

Samonas, S. and Coss, D., 2014. The CIA strikes back: Redefining confidentiality, integrity and availability in security. Journal of Information System Security, 10(3).