Intelligent Systems with Applications 28 (2025) 200607

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Enhanced Online Grooming detection employing Context Determination and Message-Level Analysis

Jake Street *, Isibor Kennedy Ihianle, Funminiyi Olajide, Ahmad Lotfi

Nottingham Trent University, 50 Shakespeare St, Nottingham, NG1 4FQ, United Kingdom

ARTICLE INFO

Keywords:
Online Grooming detection
Machine learning
Natural language processing
Context determination
Message Level Analysis

ABSTRACT

Online Grooming (OG) is a prevalent threat facing predominately children online, with groomers using deceptive methods to prey on the vulnerability of children on social media/messaging platforms. These attacks can have severe psychological and physical impacts, including a tendency towards revictimization. Current technical measures are inadequate, especially with the advent of end-to-end encryption which hampers message monitoring. Existing solutions focus on the signature analysis of child abuse media, which does not effectively address real-time OG detection. This paper proposes that OG attacks are complex, requiring the identification of specific communication patterns between adults and children alongside identifying other insights (e.g. Sexual language) to make an accurate determination. It introduces a novel approach leveraging advanced models such as BERT and RoBERTa for Message-Level Analysis and a Context Determination approach for classifying actor interactions, between adults attempting to groom children and honeypot children actors. This approach included the introduction of Actor Significance Thresholds and Message Significance Thresholds to make these determinations. The proposed method aims to enhance accuracy and robustness in detecting OG by considering the dynamic and multi-faceted nature of these attacks. Cross-dataset experiments evaluate the robustness and versatility of our approach. This paper's contributions include improved detection methodologies and the potential for application in various scenarios, addressing gaps in current literature and practices.

1. Introduction

Online Grooming (OG) attacks are well known throughout the public and within psychology-focused literature, with children being taught online safety skills to mitigate these attacks as part of the national curriculum (Department for Education, 2023). Despite the awareness of these attacks, they are still expected to be prevalent with 29% of children meeting someone they have met online who they do not know in real life (Lobe, Livingstone, Ólafsson, Vodeb, & Vodeb, 2011). The effectiveness of education as a mitigation method is not known however these attacks tend to have a significant psychological and physical impact on the child as detailed by HM Government (2021). This also includes an affinity to 'revictimisation', in which people who have experienced abuse are likely to behave in a way that victimises themselves further (Butler, Quigg, & Bellis, 2020). Throughout literature there have been many effective attempts at identifying OG attacks in a binary fashion (Borj, Raja, & Bours, 2023; Milon-Flores & Cordeiro, 2022; Villatoro-Tello, Juárez-González, Escalante, Montes-Y-Gómez, & Villaseñor-Pineda, 2012), with the use of Llama achieving accuracy scores of 1 (Nguyen, Wilson, & Dalins, 2023) on this task. However, social media platforms (SMPs) fail to implement satisfactory solutions to this problem with technical measures. In addition, the introduction of features such as end-to-end encryption on these platforms (Home Office, 2023) any potential mitigation method is even more difficult to implement, due to the inability of the platform to access messages/transcripts. Because of this, it is suggested that an endpoint solution would be required. However, as determined by Street and Olajide (2023), an endpoint text-analysis solution using Optical Character Recognition was found to be ineffective at detecting OG attacks on frequently used SMPs such as Twitter and Facebook Messenger, exacerbating the challenge of addressing these attacks. Solutions implemented to tackle the effects of OG typically focus on the signature analysis of child abuse media on SMPs (Home Office, 2023). However, this approach does not address OG attacks in real-time, and it is uncertain if end-to-end encryption will render this method ineffective. Therefore, it is essential to consider the additional complexities surrounding OG and develop solutions to overcome these limitations. It is proposed that OG attacks are more complex than other text-based social engineering attacks, as determining an OG attack from a single message is difficult due to the

* Corresponding author.
E-mail addresses: jake.street02@ntu.ac.uk (J. Street), isibor.ihianle@ntu.ac.uk (I.K. Ihianle), funminiyi.olajide@ntu.ac.uk (F. Olajide), ahmad.lotfi@ntu.ac.uk (A. Lotfi).

https://doi.org/10.1016/j.iswa.2025.200607
Received 6 December 2024; Received in revised form 22 October 2025; Accepted 8 November 2025
Available online 10 November 2025

need to meet multiple criteria, such as establishing that an adult and a child are communicating. This complexity is further compounded by the differing goals of attackers. According to Chiu, Seigfried-Spellar, and Ringenberg (2018), there are two main types of attackers: 'Fantasy Child Sex Offenders' (FCSOs), who focus on online contact and immediate sexual gratification, and 'Child Contact Sex Offenders' (CCSOs), who aim to arrange a physical meet-up with the victim and therefore the language used in these attacks is likely to differ based on these attacker types. When referring to OG attacks within literature these are predominately agreed to be text-based attacks between an adult and a child for the adult's sexual gratification.

To address this, this paper aims to determine whether an adult and a child are communicating within a transcript. With the intention to provide a greater level of detail to incident responders, as opposed to binary classification, allowing for an adjustable level of confidence in this determination to ensure that privacy is maintained as much as possible, and allow for the robust differentiation between Online Grooming and similar linguistic phenomena e.g. 'Sexting'. A clear advantage of this approach is the ability to maintain the child's privacy by a Human Reviewer having the ability to only observe messages sent by the 'Adult' actor when an Adult-Child context has been established. However a disadvantage of this proposed approach is that using Context Determination as a standalone system other Adult-Child contexts are likely to be identified. Therefore additional insights are required for real world implementation. This topic has been explored in the context of author profiling determination, which is common in literature (Argamon, Koppel, Pennebaker, & Schler, 2009; Goswami, Sarkar, & Rustagi, 2009; Nguyen, Gravel, Trieschnigg, & Meder, 2021; Peersman, Daelemans, & Vaerenbergh, 2011). However, many methodologies use blogging websites for author profiling due to their vast datasets and labelled data, rather than focusing on the context of OG (OG). While author profiling has been somewhat considered in the context of OG within the PAN13 competition (Rangel & Rosso, 2013), the robustness of these approaches needs further application to OG detection. Robustness is a key consideration that needs to be made towards any OG solution. In the context of this paper, there are limitations with datasets currently available due to the language used in these transcripts likely being outdated in terms of their 'textspeak' as well as the topics of conversation that are discussed. In addition, there are likely changes in the method of initiation that is used by the attacker due to the decrease in usage of 'chatroom-based' SMPs and the increase in 'reel-based' SMPs such as TikTok and Instagram (Office of Communications, 2023). Current messaging on these new-age SMPs contain features/data that were not previously available; replies to a specific message, the sending of media in conversation & the ability to reply to media in chat, destructive messages/media, and video calls. Whereas focusing on text-based messaging alone is likely to have real world application, it is possible that the presence additional features within a given transcript provides a greater deal of information that will allow for a greater accuracy of determinations. However to ensure generalisability, real world solutions need to consider these additional features and how these may play a role in these attacks. To date, no method has utilised identifying the Actor–Actor dynamic to classify OG, nor has there been an approach using Actor Significance as a concept to determine OG. In addition, the polarisation of message determinations has not been addressed within literature. This paper shall describe this Actor–Actor context determination approach, with the specific contributions of this study being:

1. A novel Message-Level Analysis leveraging advanced models such as BERT and RoBERTa specifically in the context of OG detection.
2. A novel Context Determination approach for classifying actors and their interactions, which has the potential to allow generalisation to group chat settings.
3. Introduction of Actor Significance Thresholds and Message Significance Thresholds to better address the specific challenges of OG detection.
4. Cross-dataset experiments to evaluate the robustness of the proposed context determination approaches and message-level analysis. Furthermore, exploring its potential applicability in different scenarios, demonstrating its versatility and robustness for OG detection.

This paper is organised into several sections. Section 2, an overview of related work will be presented, providing a comprehensive understanding of the current technical and psychological perspectives on OG attacks. Section 3, the Proposed Methodology section will then offer a theoretical framework for the investigation. This will be followed by the Experimental Methodology Section 4, detailing the configurations and processing methods used, Section 5 presents the results and the discussion allowing for a thorough analysis of the findings. Finally, the Conclusion summarises the outcomes and suggests directions for future research in Section 6.

2. Related work

Initial research into OG attacks primarily focused on identifying the linguistics and psychology of groomers to develop frameworks for understanding these attacks. One of the most prominent frameworks was proposed by O'Connell (2003), which theorised distinct 'phases' of OG, with the severe sexual phases typically occurring towards the end of the communication. This was validated by Kontostathis, Edwards, and Leatherman (2010), who used K-means clustering on chat logs from Perverted Justice (PJ)[1] a vigilante justice website that publishes transcripts of honeypot interactions with offenders. This approach supported the rule-based method known as 'chatcoder', categorising transcript characteristics into eight categories. Due to the lack of negative OG transcripts within the PJ dataset, it is difficult to validate the real-world application of any approach claiming accuracy based solely on PJ data. Therefore, the PAN12 dataset (Inches & Crestani, 2012) was employed to broaden the scope and enhance the classification challenge of OG. Most approaches treat this as a typical binary classification problem, employing various text preprocessing methods and training models with different configurations to optimise performance, typically evaluated using F-score metrics across diverse use cases; Identifying Individual Predatory Messages using $\beta = 3$ with a focus on reducing false negatives (Inches & Crestani, 2012), and Identifying Predators using $\beta = 0.5$ with more of a focus on reducing false positives (Inches & Crestani, 2012). Regarding datasets used in the literature, the focus has primarily been on two main datasets. While other studies have explored OG detection methods in different contexts (Cheong, Jensen, Gudnadottir, Bae, & Togelius, 2015; Chiang & Grant, 2019; Kloess, Seymour-Smith, Hamilton-Giachritsis, Long, Shipley, & Beech, 2015), these investigations do not specifically evaluate the robustness of a potentially generalised OG solution.

Various preprocessing and encoding methods are employed within OG classification tasks, including Word2Vec (Borj et al., 2023; Ebrahimi, Suen, & Ormandjieva, 2016), Bag of Words (Ebrahimi et al., 2016; Villatoro-Tello et al., 2012), and Term Frequency Inverse Document Frequency (TF-IDF) (Milon-Flores & Cordeiro, 2022; Villatoro-Tello et al., 2012). These methods are integral to optimising text data for effective analysis and classification. In the context of OG detection, preprocessing methods such as stopword removal (Cheong et al., 2015) and repeated fixed letter removal (Cheong et al., 2015) have been utilised. However, the efficacy of these methods has been questioned, particularly regarding common NLP preprocessing techniques like punctuation removal, which may inadvertently remove crucial

[1] perverted-justice.com

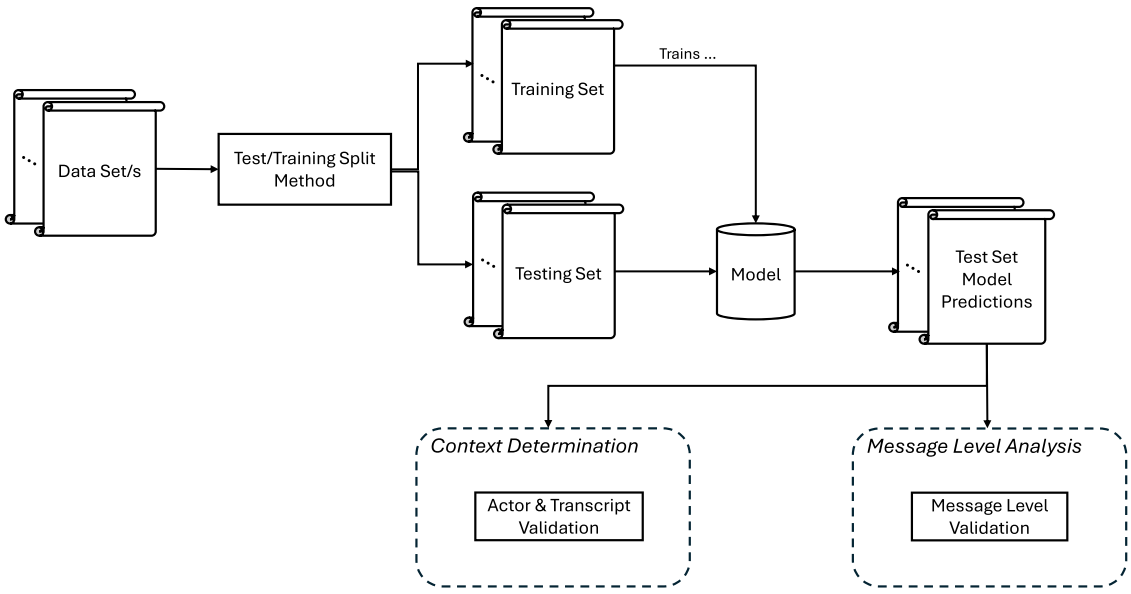

Fig. 1. An overview of the methodology depicting the testing methods for Context Determination and Message Level Analysis.

information used in OG classification (Villatoro-Tello et al., 2012). For classification tasks on the PAN12 dataset, a range of models have been explored, including traditional approaches such as Support Vector Machines (SVM) (Inches & Crestani, 2012), Naive Bayes (NB) (Inches & Crestani, 2012; Michalopoulos & Mavridis, 2011), Logistic Regression (Inches & Crestani, 2012; Rangel & Rosso, 2013), Neural Networks (Ebrahimi et al., 2016; Villatoro-Tello et al., 2012), and transformer-based models like BERT/RoBERTa (Borj et al., 2023) and Llama (Hamm, 2025; Nguyen et al., 2023). Recent literature on the PJ dataset has primarily focused on clustering methods (Chiu et al., 2018; Kontostathis et al., 2010), owing to its 'all positive' nature where all interactions are assumed to involve grooming. These clustering approaches have identified distinct phases within transcripts, aligning with frameworks proposed by previous studies (Edwards et al., 2009; O'Connell, 2003).

3. Methodology

The critical factor in differentiating OG interactions from other types of interactions lies in reliably determining the context of the interaction. Specifically, this involves establishing whether an adult is interacting with a child and if this interaction is conducted in a grooming manner. Once this context is established, additional indicators such as the use of sexual language can further identify an interaction as OG. Although user profile attributes could provide this information, this method is vulnerable as both attackers and children often falsify their ages on these platforms.

This paper follows two primary methodological steps, as shown in Fig. 1, - Message-Level Analysis (MLA) for which each phrase within a transcript is scored using a variety of models, including BERT, RoBERTa, SVM, and Naive Bayes (NB); Context Determination - Following the MLA, a comprehensive analysis of each transcript is performed to determine the Full Transcript Context. This process involves evaluating the context at the transcript level for each actor to establish whether the interaction involves both an adult and a child. If such a determination can be made, the transcript is labelled accordingly. Transcripts that clearly involve both an adult actor and a child actor are labelled as OG interactions. The accuracy of this labelling is then analysed. It is important to note that while this approach may be valid within the tested datasets, it may not directly transfer to real-world contexts without additional corroborative insights.

The research questions intended to be answered from this are listed below.

1. Can a robust model be formed from taking a full transcript context determination approach?
2. What impact do message determination thresholds (t values) and AST values have on the transcript OG determination accuracy?
3. What model is recommended to be used in different use cases?

The key piece of information that would help to differentiate an interaction as OG as opposed to other interactions is being able to reliably determine the 'context' of the interaction i.e. Is an adult interacting with a child, and is this interaction in a 'grooming way'? Once this piece of information is established the presence of other indicators, e.g. Sexual Language, can help to establish a given interaction as OG. This information could be obtained through attributes on the user's profile however this approach is likely to be easily mitigated with knowledge of this method being made public as well as both attackers and children commonly lying about their age on these platforms.

3.1. Message level analysis

The Message Level Analysis (MLA) process involves two main approaches as shown within Fig. 2 'Inter-set Training', in which testing and training occur on the same set of texts which indicates the accuracy of OG detection within a well defined OG 'use case' and 'Cross-set Training', in which testing and training occur on two different sets of text which allows observation to be made as to the robustness of a system across different OG use cases. In the context of a real-world implementation the inter-set experiments provide some indication as to how this would perform on the chatroom attack vector and the cross-set experiments provide some insight into how this will perform at detecting current-day attacks on prevalent SMPs.

As depicted in the 'Results Metrics' section of Fig. 2, four metrics are proposed to evaluate the performance of each tested approach. These metrics include True Positive (TP), representing correct 'A' or 'C' determinations; Incorrect Child Determination (IC), indicating falsely identified 'child messages'; Incorrect Adult Determination (IA), denoting falsely identified 'adult messages'; and Omissions (O), which counts messages excluded from the test set due to insufficient polarisation towards either adult or child determination based on the Message Threshold (t) value used. The MLA process focuses solely on classifying messages as either 'Adult' or 'Child'. It is crucial to consider the types of texts included in this classification, as including non-OG texts may skew message determination values, potentially increasing the False

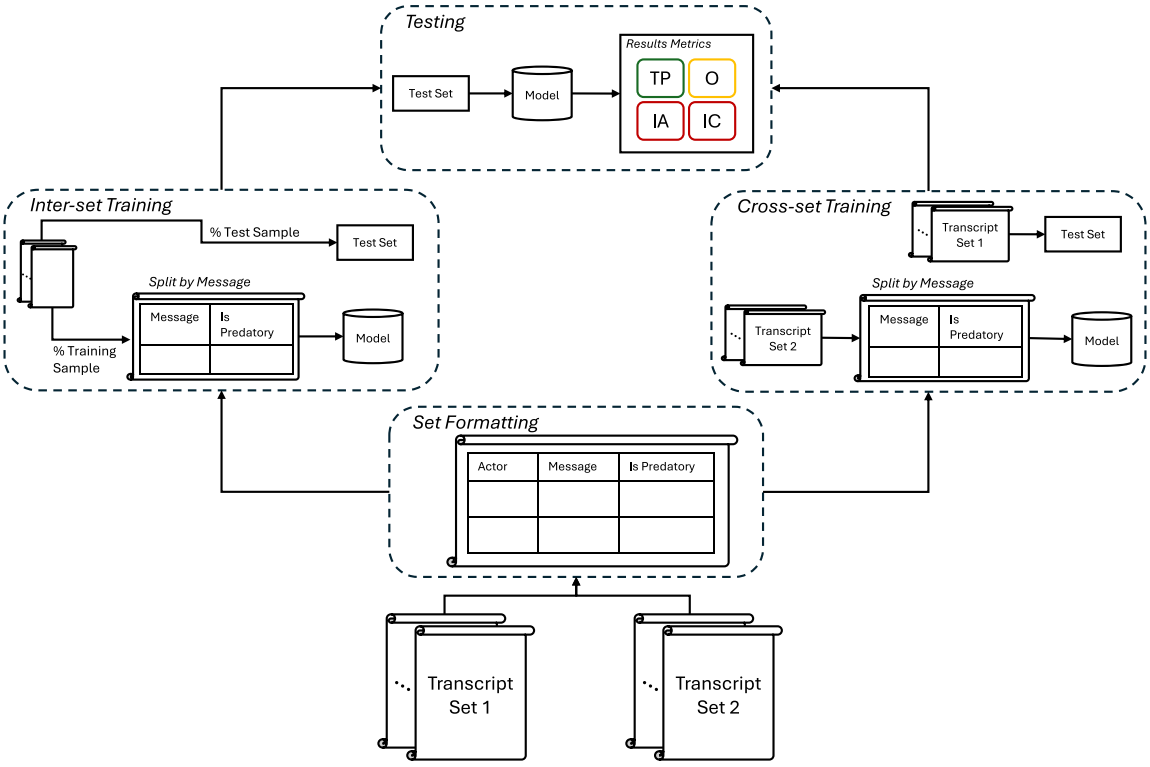

Fig. 2. Message Level Analysis (MLA) process showing Inter-Set and Cross-Set approach.

Positive rate in the Context Determination process and minimising the impact of Omissions on the dataset. This assumption requires empirical validation. Therefore, this paper will exclusively analyse OG-positive transcripts within the MLA process, acknowledging the inability to determine whether messages in negative OG transcripts from PAN12 are classified as 'Adult' or 'Child'.

3.2. Context determination

Context determination aims to identify if an adult and a child are communicating with one another within a transcript. This is defined as a transcript having an Adult-Child (AC) context. While this information can be obtained within the sign-up process by using the date of birth, in a real-world context both the attacker and/or the child can lie about their true age to bypass any mitigation method that is dependent on this information with ease. However, by implementing a mitigation that focuses on language usage to determine age there is difficulty in an attacker mitigating this. Nevertheless, it should be noted that due to the variety of methods used by attackers, attackers can simulate being a child within their attack (O'Connell, 2003) and the language used is likely to reflect this. While this method would be out of the scope of this investigation, it is suggested that implementation to observe 'language context differences' to identify if this method is being used, i.e. an actor context changing to an adult when the sexual stage (O'Connell, 2003), could be used to identify this mitigation method. Therefore, it is theorised that the development of a method that can determine the context of a given interaction between two actors would provide a key piece of evidence in determining the complex concept of OG. In determining this context each of the two actors can be determined to be one of the following categories;

- Significantly an Adult (A)
- Significantly a Child (C)
- Not Significantly an Adult or Child (NS)

The 'significance' element in this determination refers to the dynamic threshold used in the Context Determination process to classify messages as Adult or Child. This threshold allows for adjustable confidence levels, which can be tailored to observe the optimal confidence needed for specific SMP use cases. An Actor Significance P value threshold (AST) is employed, representing the probability that a given determination, along with an equally or less likely determination, could occur by random chance. Different AST values will be tested to evaluate their impact on the occurrence of false positives and false negatives. It is anticipated that higher AST values, indicating less significance, will result in a lower or equal frequency of false negatives but a higher or equal frequency of false positives. Conversely, lower AST values, indicating greater significance, are expected to lead to a higher or equal frequency of false negatives and a lower or equal frequency of false positives. In the context of SMPs, minimising the false positive frequency in OG determinations is typically preferable. However, it is important to note that in this investigation, the Adult/Child context insight is part of a broader determination of OG, expected to be combined with other insights. Therefore, a lower false negative frequency is likely preferable to allow other insights to refute a transcript as indicative of OG. These individual actor determinations can be mapped to an overall transcript context based upon the rule of combining the contexts unless one or both are deemed 'Non Significant', in which case the full context determination is deemed Non Significant. The methodology that shall be followed when using this approach is described within Fig. 3 which demonstrates the three main processes in determining a given transcript's context. These are Message Splitting and Omission, in which each message is attributed to the actor that sent it while also omitting messages that are not significant enough based upon the t value; Determining if each actor is an Adult or Child and the significance of this, where the AST experiment value is used to determine the threshold of significance; and Determine Transcript Context, in which the determinations from both actors are brought together to give an overall transcript context.

This investigation focuses exclusively on peer-to-peer communications, where transcripts involve interactions between two unique

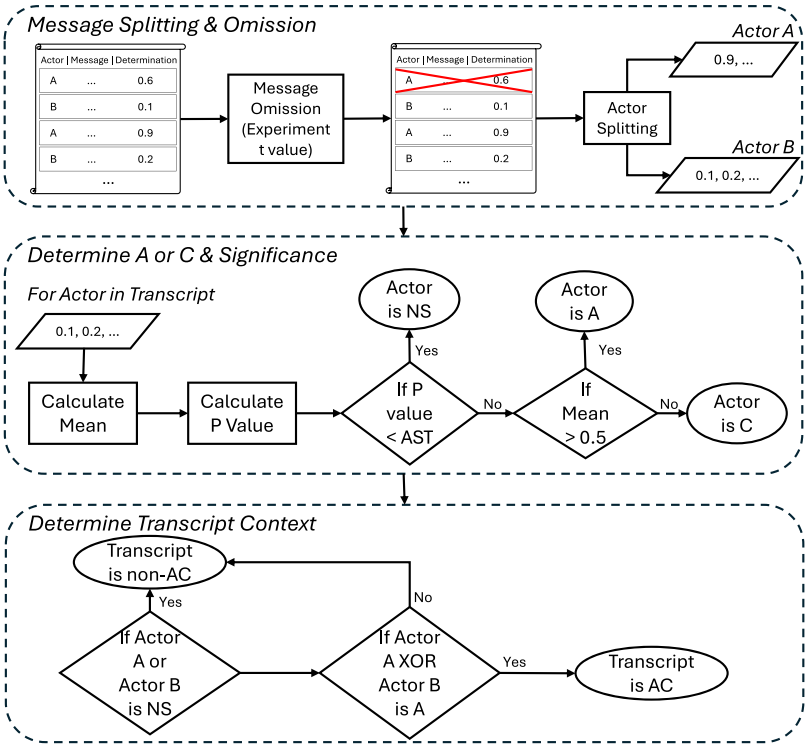

Fig. 3. Context determination process with AST and t value application.

actors. While this approach raises some privacy concerns of the individuals using the platform to some degree, and with these interactions being of a sensitive nature users of all types would likely reject a system of this nature due to privacy concerns. Therefore the motivation of this work is intended to observe if the involvement of human reviewers, as seen with harmful content moderation, could be reduced in a potential implementation. By being able to reliably obtain the insight that an adult and a child are conversing, it is suggested that when this is applied alongside sexual language detection then a complex OG determination can be made which is likely to be robust against other messaging that is similar in language to OG e.g. 'Sexting', defined as the consensual sending of sexual messages/media. As well as the proposed approach using AST value to determine the degree of confidence that is held for an Actor being an Adult or a Child, the same ideology can be applied to the individual messages that make up this determination. By using a 'Message Threshold Value' messages that are of low confidence in determining an Adult or a Child can be omitted from consideration. This provides additional contextualisation to the problem of OG as not every message that is sent between two actors could be used to determine their age. A key point to note with the Message Threshold Value is that the frequency of low-confidence messages linked to an actor does not affect the overall actor determination process as these are simply omitted from consideration. This was an intentional choice as it is to be expected that both Adult and Child actors will send messages that do not significantly link to being either an Adult or a Child. Table 1 shows an example transcript within the dataset with both low-confidence and high-confidence messages. As shown the first two messages display high Adult/Child confidence determinations i.e. it can easily be determined which actor is an adult/child, whereas, with the other two messages, there is a lack of confidence as it is likely that both an adult or a child could send the message. The 'Prediction' values shown within Table 1 were made from the BERT model.

Message Threshold Values, t, from 0.2-0.45 shall be analysed within this investigation, this value will be applied to determine if a given message is to be considered within the Actor context determination. With a determination of '0.5' being the lowest confidence message

Table 1
Transcript confidence example within PAN12 (data frame format).

	Authors	OG	Message	Prediction
5	1	1	im old 49 here old enough to be your daddy	0.99998
6	0	0	um my dad's older	0.01687
19	1	1	r u there	0.38972
20	1	1	u have a good day	0.32781

determination value that could be given. A message will be omitted from consideration if Eq. (1) is true, where n is the determination value of the given message.

$$(0.5 + t) > n > (0.5 - t) \tag{1}$$

Only four set t values shall be used within this investigation. However, it is expected that the optimum t value could be obtained from regression if these t values are found to have a significant effect on the overall Context Determination. The purpose of the Full Transcript Context Determination is to evaluate the validity of using an insight such as Adult/Child context interaction to determine OG, with future insights. All messages within a transcript will be utilised in this determination, therefore a message towards the beginning of the communication will have the same weighting in this analysis as one towards the end. Where t is the message threshold used to be included within the determination (only applicable to BERT and RoBERTa). As depicted in Fig. 3 Actor A and Actor B reference two unique actors within a peer-to-peer communication, Actors are then determined to either be 'A', significantly an Adult; 'C' significantly a Child; or 'NS' not significantly either. Based upon that a check is made to see if there is an Adult Child combination using the following logic as shown within the 'Determine Transcript Context' process described in Fig. 3: (Actor A is Adult XOR Actor B is Adult) AND (Actor A is Child XOR Actor B is Child).

The source code for both the Message Level Analysis and Context Determination approaches can be found here (Street, 2025).

4. Experimental methodology

This section shall give an overview of the implementation choices made as part of this investigation, as well as highlighting challenges faced within the preparation of the datasets.

4.1. Datasets

Due to the ethical and legal considerations that must be made within obtaining OG transcripts, there is a scarcity of data that can be used. The majority of literature investigating OG either uses the PAN12 dataset or a version of the Perverted Justice dataset including transcripts from 2006 up to 2016. However, most investigations only use a limited portion of the available data e.g. in the investigation by Kontostathis et al. (2010). This paper uses all available transcripts from the PJ dataset[2] (at the time of publication) which have followed a semi-automated labelling process by the research team. The transcripts were published using different HTML formatting over time the use of HTML parsing allowed for the majority of transcripts to be automatically labelled, as predators were identified using different styling or the predator's username was in the web page title. Nevertheless, some transcripts required manual labelling from the research team. There have been investigations, of both a psychological and technical nature, using other datasets that have not been made publicly available due to these aforementioned considerations (Cheong et al., 2015; Chiang & Grant, 2019; Kloess et al., 2015). Therefore within this investigation, the two datasets shall be used - from PJ and the PAN12 dataset[3](Inches & Crestani, 2012). In contrast to existing OG detection literature, this study utilises the 'full version' of the PJ dataset, which includes all available data up to the present, unlike previous studies that could only access datasets available at the time of their publication. It is important to note that 80% of an earlier iteration of the PJ dataset (spanning 2006–2012) is included in the PAN12 dataset (Inches & Crestani, 2012), alongside transcripts from three other sources. The PJ dataset distinguishes itself by labelling each transcript with a groomer and a victim for every message, facilitating the classification of each message as either Adult or Child based on these designations. However, in the PAN12 dataset, such labelling is absent for 'negative OG' transcripts, which hinders the ability to contextualise these interactions. Nevertheless, the same methodology is employed in PJ can be applied to 'positive OG' transcripts within PAN12. Since the PJ dataset consists of .html webpages from the host website, a preprocessing and labelling step is necessary to prepare them for analysis. Each webpage contains a transcript with a dynamic preamble, necessitating HTML parsing to extract messages and essential information for identifying the attacker and victim in each transcript. This process involves examining the file name, typically indicating the attacker, although, in some instances, attackers may use multiple usernames, rendering this approach unreliable. Each transcript must undergo careful analysis to determine its formatting requirements, ensuring the correct processing steps are applied. Algorithm 1 outlines this

As shown by Algorithm 1, a known attacker's file is used, this allows for the semi-automated labelling of messages which ensures that each message does not have to be labelled, and instead, the attacker name is searched for amongst the message lines. These attackers are entered into this file when the 'attacker entry' process, as shown in Algorithm 1, is run. Following this each transcript is then checked for the 'blueBold' formatting, which purely looks at this to determine attacker message lines and if this is not present name processing is used which utilises the ' 'knownAttackers' and HTML file titles to attempt to label the attacker lines. To ensure that there are no errors within this a check

Algorithm 1 'peverted-justice.com' (PJ) webpage processing.

```
knownAttackers ← AttackersFile
for htmlPage do
    allText ← htmlPage.read
    attackerTitle ← htmlPage.replace(fileExtension)
    if blueBold ∈ allText then
        FormattedProcessing
    else if attackerTitle ∈ allText | attackerTitle ∈ knownAttackers then
        NameProcessing
    else
        AttackerEntry
    end if
end for
```

```
<span class='blueBold'>DavieWants2</span> [1:58 PM]: sorry i just feel that YOu can have ahappy life, just dont do anything stupid, lie YOu siad ther are lots of sicko out ther kid, so becareful <br />

SnapShotDeath [1:59 PM]: no i swear i do what u say <br />

SnapShotDeath [1:59 PM]: u said u buy me stuff <br />

<span class='blueBold'>DavieWants2</span> [2:00 PM]: but kid, i want YOu to be happy too, so tread carefully itd dangerous out three boi <br />

SnapShotDeath [2:00 PM]: u said u make me happy <br />
```

Fig. 4. Example of a transcript for Formatted Processing (user 'DavieWants2' from PJ).

```
7:13:54 PM armysgt1961: CAN I GET YOUR NUMBER <br />

7:14:12 PM peekaboo1293: im not suposed 2 give my number if i dont kno u yet<br />

7:14:17 PM armysgt1961: OH <br />

7:14:27 PM peekaboo1293: i haft to call dad with my calling card so mom dont kno<br />

7:14:37 PM peekaboo1293: so dad sends me calling cardz hehehe<br />

7:14:50 PM armysgt1961:I CAN BUY U SOME MINS <br />
```

Fig. 5. Example of a transcript for Name Processing (user 'ArmySgt1961' from PJ).

```
8/1/2006 7:22:23 PM: {superboy_93}just hanging out<br />

8/1/2006 7:22:42 PM: {adams217}good deal...how was ur day<br />

8/1/2006 7:25:31 PM: {superboy_93}cool went sk8ng n swimming<br />

8/1/2006 7:25:40 PM: {superboy_93}sorry some dood was bugging me<br />

8/1/2006 7:25:54 PM: {adams217}no probblem<br />

8/1/2006 7:25:59 PM: {adams217}where 'd u go swimming<br />
```

Fig. 6. Example of a transcript for Attacker Entry (user 'Adamou217' from PJ).

on the number of messages for both labels is considered, and if there is a significant difference (more than 70% of one label) between the labelling frequency within a transcript this is then manually checked and 'attacker entry' process is then used to make adjustments.

As shown by the webpage example in Fig. 4 there is 'blueBold' formatting on the message lines that have been sent by the attacker, but this formatting is not present for the honeypot message lines. As shown by Fig. 5 there is an exact match on the message line of the attacker's username with the messages they have sent, however as seen within Fig. 6 a different username is being used by the attacker 'adams217' compared to the title of the webpage 'Adamou217' therefore attacker entry is needed to include 'adams217' to the list of attackers. In addition, different formatting options were used on these webpages there were inconsistent prefixes to the message lines on these webpages. For example, some message lines include the time, some

[2] perverted-justice.com

[3] https://pan.webis.de/clef12/pan12-web/sexual-predator-identification.html

include 'AM/PM', some include different brackets for formatting, and some include comments on specific messages that were not included within the communication. As seen across Figs. 4, 5, and 6 there are differences in this time formatting with some including the full date, and some including the time to the resolution of seconds.

Verification steps to ensure that labelling has been done correctly has been conducted as part of this investigation. Primarily from the running of the context determination process if a transcript did not have two separate actors this would cause errors when running context determination which could then be rectified. In addition, consecutive actor verification occurred in which instances of 10 message lines in a row from the same actor were investigated. While these issues did not occur, there is potential for small sections of the transcripts to be mislabelled if different actor names had been used in the same transcript.

4.2. Preprocessing methods

Similar to the investigation by Villatoro-Tello et al. (2012) preprocessing methods will not be used in our approach, beyond the exclusion of transcripts shorter than 10 messages in length for the PAN12 dataset and the exclusion of email transcripts within the PJ dataset used in this investigation. It should be noted that, while this may improve the accuracy of determinations, this is unlikely to be transferable/robust to a new age real-world context due to the move away from the use of emoticons formed of punctuation marks, e.g. '(:-*)' for kiss (Villatoro-Tello et al., 2012), that are likely being used within these OG determination methods.

4.3. Models

Naive Bayes (NB) and Support Vector Machine (SVM) have commonly been seen implemented as solutions to the binary classification of OG. In terms of the technical configurations used within this investigation, for NB the 'sklearn.naive_bayes' module was used (Pedregosa et al., 2011) alongside the feature extraction module as part of this module. For the SVM model, this same module was used alongside the 'pipeline' with a 'countVectoriser' approach (Pedregosa et al., 2011). For both the transformer-based models (BERT and RoBERTa) 'tensorflow' was utilised (Abadi et al., 2015). For RoBERTa the 'RoBERTa-base' tokeniser was used (Liu et al., 2019), whereas for BERT the BERT tokeniser within tensorflow was used (Abadi et al., 2015).

5. Results and discussion

The results within this investigation are split into two parts Message Level Analysis and Context Determination. The purpose of these two sections is to first, within Message Level Analysis, establish the baseline accuracy of models in both an inter-set and cross-set context, as well as to observe the impact of t values. Then the second section, Context Determination, allows for a comparison to be made against this established baseline to put this in the real-world context of OG detection, as well as observing the impact that AST values have on this.

5.1. Message level analysis results

Across the PJ sample test set that was trained on the remaining 80% of the PJ dataset, the MLA results are shown within Table 2. It should be remembered that despite the size of the PAN12 dataset in comparison to PJ, all messages within negative OG transcripts are excluded from MLA. Therefore in comparing the performances across both test sets this needs to be taken into account, as well as the increased training size for the models. Within Table 2 'TP' refers to True Positive, 'IA' refers to Incorrect Adult Determination, 'IC' refers to Incorrect Child determination, and 'O' refers to Omissions. Note that the percentage values do not sum to 100% as the Omissions percentage value refers to the total percentage of messages that were omitted from consideration whereas the other percentage values refer to the respective metric of the messages that were still included within the process. These omitted messages consist of lower Adult-Child polarised messages as can be seen in Table 1. A total of 219,168 messages were included within this MLA analysis over the PJ dataset.

Table 2
MLA results across PJ sample.

Model	TP	IA	IC	O
SVM	76.0%	13.5%	10.5%	0%
NB	74.7%	12.0%	13.3%	0%
BERT (t=0.2)	82.2%	9.3%	8.5%	9.2%
BERT (t=0.3)	83.7%	8.4%	7.7%	15.3%
BERT (t=0.4)	85.6%	7.8%	6.6%	22.7%
BERT (t=0.45)	87.3%	7.1%	5.6%	30.0%
RoBERTa (t=0.2)	83.3%	7.7%	9.0%	8.8%
RoBERTa (t=0.3)	84.9%	6.9%	8.2%	14.7%
RoBERTa (t=0.4)	87.2%	6.0%	8.9%	23.6%
RoBERTa (t=0.45)	88.8%	5.3%	5.9%	30.4%

Table 3
MLA results across PAN12 (positive OG transcripts).

Model	TP	IA	IC	O
SVM	63.4%	25.5%	11.2%	0.0%
NB	61.4%	23.4%	15.2%	0.0%
BERT (t=0.2)	64.3%	25.8%	9.9%	7.3%
BERT (t=0.3)	64.8%	26.0%	9.2%	11.1%
BERT (t=0.4)	65.6%	26.3%	8.1%	16.3%
BERT (t=0.45)	66.4%	26.7%	6.9%	20.8%
RoBERTa (t=0.2)	65.1%	26.6%	8.3%	7.5%
RoBERTa (t=0.3)	65.7%	26.8%	7.5%	12.2%
RoBERTa (t=0.4)	66.5%	26.9%	6.6%	20.2%
RoBERTa (t=0.45)	67.2%	27.0%	5.8%	26.3%

Table 3 shows the MLA results from the PAN12 dataset on transcripts that contain more than 30 messages and contain messages deemed to be predatory (assumed to be Adult), A total of 702,419 messages were included within this MLA analysis over the PAN12 dataset.

The results from the MLA on PAN12 as shown within Table 3 initially suggest that these models are not robust in determining Adult /Child across datasets, weighted significantly to incorrect Adult determinations. Following these results, the weightings between both labels in the PAN12 dataset were analysed and it was found that 68% of the messages were labelled as predatory. This value seems unlikely to be a fair reflection of the transcripts in terms of Adult/Child interaction due to how great the difference is when considering the vast size of the dataset.

When comparing the MLA results across PAN12 and PJ there is a significant difference in results which leads to the initial conclusion that cross-dataset determinations lack robustness. Despite this difference, it is theorised that when combining these determinations as part of full context determination may allow for an increase in overall accuracy. An interesting observation to be made is the True Positive Accuracy of SVM without any omissions. It is possible that if there were no omissions (i.e. a t value of 0 is used) within the transformer-based models, SVM in this instance may outperform these models. Another important thing to note is the skew towards Adult determinations, as seen with the Incorrect Adult determination metric, within the PAN12 dataset. The mean difference between the sum of Incorrect Adult and Incorrect Child determinations within the PJ sample set was −0.02% whereas the mean difference for PAN12 between Incorrect Adult and Incorrect Child was +17.4%. It is suggested that to counteract the 'skew' that is apparent when conducting this cross-dataset analysis, dynamic t values could be used for both the Adult and Child thresholds for omission. In this case, an inflated t value for the Adult determinations (determinations more

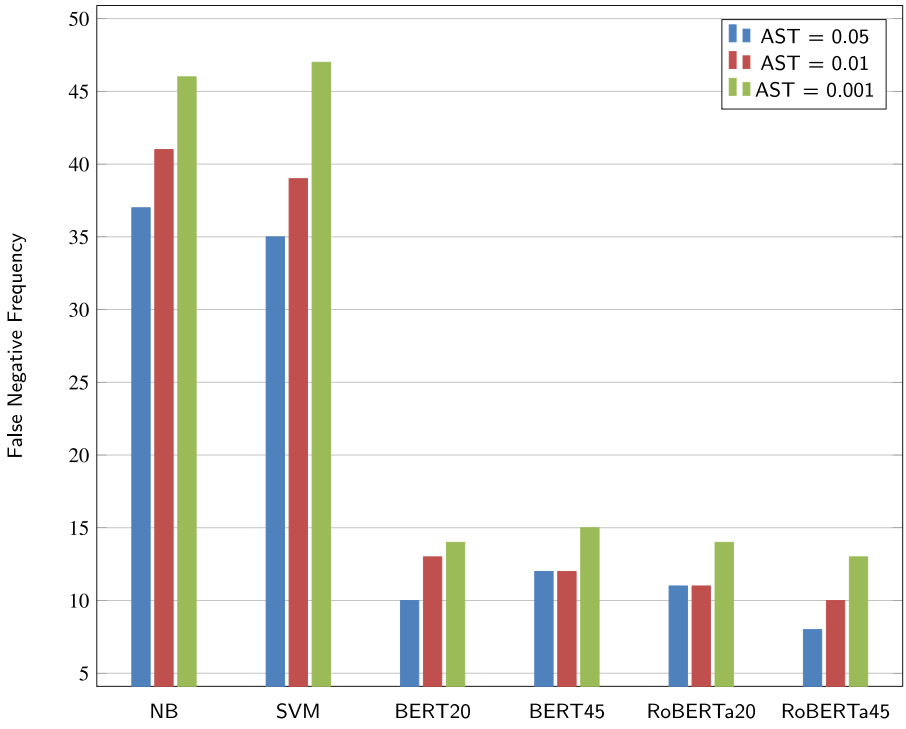

Fig. 7. PJ sample full context determination FN frequencies.

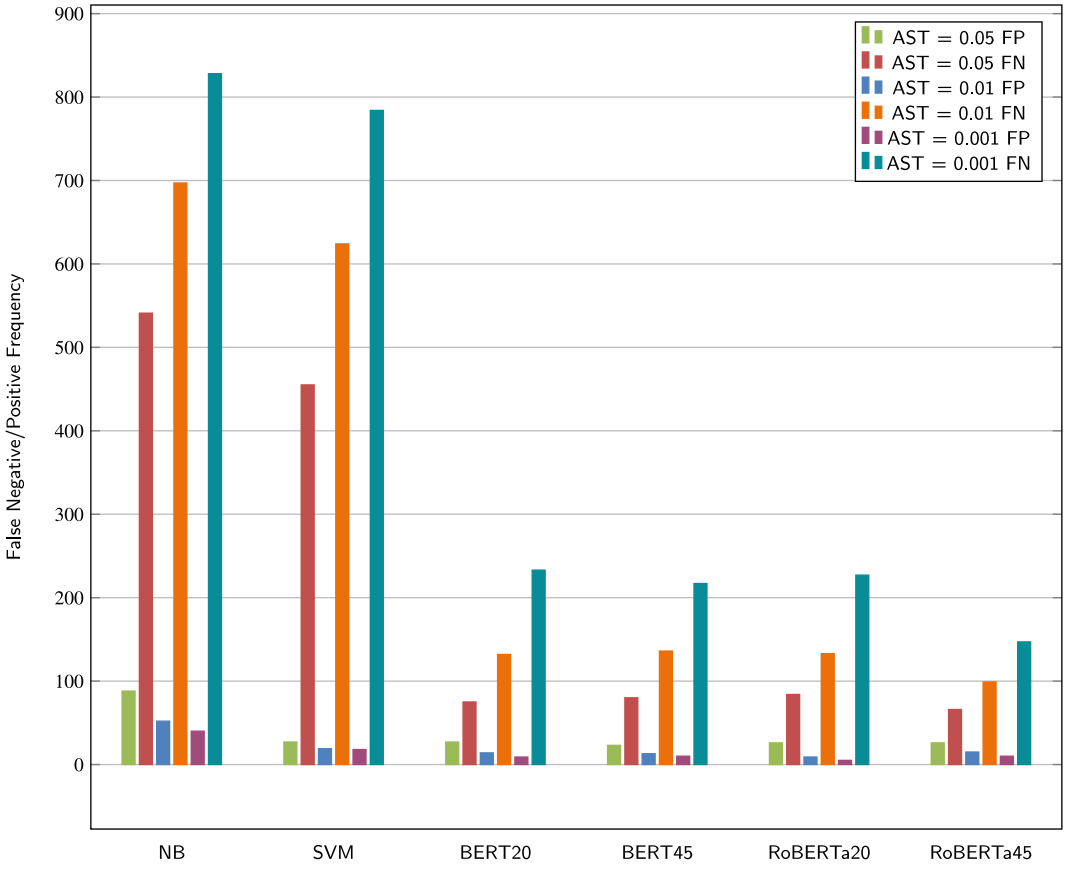

Fig. 8. PAN12 full context determination FP & FN frequencies.

Table 4
PJ sample full transcript context determination with an AST value of 0.05.

Model	TP	FN	F1
SVM	118	35	0.871
NB	116	37	0.862
BERT (t=0.2)	143	10	0.966
BERT (t=0.3)	141	12	0.959
BERT (t=0.4)	141	12	0.959
BERT (t=0.45)	141	12	0.959
RoBERTa (t=0.2)	142	11	0.963
RoBERTa (t=0.3)	143	10	0.966
RoBERTa (t=0.4)	144	9	0.970
RoBERTa (t=0.45)	145	8	0.973

Table 5
PJ sample full transcript context determination with an AST value of 0.01.

Model	TP	FN	F1
SVM	114	39	0.854
NB	112	41	0.845
BERT (t=0.2)	140	13	0.956
BERT (t=0.3)	140	13	0.956
BERT (t=0.4)	140	13	0.956
BERT (t=0.45)	141	12	0.959
RoBERTa (t=0.2)	142	11	0.963
RoBERTa (t=0.3)	142	11	0.963
RoBERTa (t=0.4)	143	10	0.966
RoBERTa (t=0.45)	143	10	0.966

Table 6
PJ sample full transcript context determination with an AST value of 0.001.

Model	TP	FN	F1
SVM	106	47	0.819
NB	107	46	0.823
BERT (t=0.2)	139	14	0.952
BERT (t=0.3)	139	14	0.952
BERT (t=0.4)	138	15	0.948
BERT (t=0.45)	138	15	0.948
RoBERTa (t=0.2)	139	14	0.952
RoBERTa (t=0.3)	139	14	0.952
RoBERTa (t=0.4)	139	14	0.952
RoBERTa (t=0.45)	140	13	0.956

than 0.5) could be used with the expectation to reduce this 'skew'. This is based on the assumption that the Incorrect Adult Determinations are closer in value to the omission threshold than the True Positive adult determinations. A further investigation is required to validate if this t value adjustment method, and assumptions about the Incorrect Adult Determinations, have validity.

5.2. Context determination results

5.2.1. PJ results

The below results are for PJ 20% with a total of 152 transcripts which all have a context of 'AC', therefore only the True Positive and False Negative metrics are used. Tables 4–6 gives an overview of the results obtained.

Fig. 7 shows the False Negative frequencies amongst different models and message threshold values (t). These t values are denoted in both Figs. 7 and 8 within the model labels that have the suffix '−20', denoting a t value of 0.2, and '−45', denoting a t value of 0.45. As described the lower the Actor Significance Threshold (AST) value the reduction in False Negatives. It is anticipated that there will be a noticeable decrease in False Negatives when the message threshold value is increased for RoBERTa. Conversely, the opposite effect seems to occur when this approach is applied to BERT, as evidenced by the Actor Significance p-value of 0.05. It is difficult to make conclusions on the application of these results without observing the False Positive rate of these models, which is unable to be done within the PJ sample as all transcripts are OG positive. This shall be investigated within the analysis of PAN12.

Table 7
PAN12 full transcript context determination with an AST values of 0.05, 0.01, and 0.001.

Model	TP	FP	TN	FN	F1
SVM	717	27	10309	455	0.748
	548	19	10317	624	0.630
	388	18	10318	784	0.492
NB	631	88	10248	541	0.670
	475	52	10284	697	0.559
	344	40	10296	828	0.442
BERT (t=0.2)	1097	27	10309	75	0.956
	1040	14	10322	132	0.934
	939	9	10327	233	0.886
BERT (t=0.3)	1098	28	10308	74	0.956
	1044	14	10322	128	0.936
	947	10	10326	225	0.890
BERT (t=0.4)	1095	28	10308	77	0.954
	1042	14	10322	130	0.935
	965	11	10325	207	0.899
BERT (t=0.45)	1092	23	10313	80	0.955
	1036	13	10323	136	0.933
	955	10	10326	217	0.894
RoBERTa (t=0.2)	1088	26	10310	84	0.952
	1039	9	10327	133	0.936
	945	5	10331	227	0.891
RoBERTa (t=0.3)	1086	26	10310	86	0.951
	1044	10	10326	128	0.938
	962	6	10330	210	0.899
RoBERTa (t=0.4)	1107	27	10309	65	0.960
	1076	9	10327	96	0.953
	1018	7	10329	154	0.927
RoBERTa (t=0.45)	1106	26	10310	66	0.960
	1073	15	10321	99	0.955
	1025	1E0	10326	147	0.929

5.2.2. PAN12 results

These results differ in terms of context when comparing these to the full context results for PJ where each transcript within PJ is classed as AC, this is not the case for PAN12 with the transcripts being any possible context. It should be noted that even OG negative transcripts could correctly be determined to be AC if the interaction is between an Adult and a Child but is not of a grooming nature. This is where further insight analysis, such as a sexual severity insight, would factor in to make a complete OG determination. Table 7 shows the results from using an AST value of 0.05, emphasising reducing False Negatives (FN); 0.01; and 0.001, emphasising reducing False Positives (FP).

Fig. 8 demonstrates the False Positives and False Negatives within the PAN12 dataset when using different models and t values. From this it can be suggested that a greater t value leads to greater accuracy of results, therefore greater weight should be made towards the key polarising tokens/messages that indicate either an adult or a child. In addition, it can be suggested that, as the False Negative frequencies are significantly greater than the False Positive frequencies amongst all models & t values, a less significant Actor Significance value could be implemented to observe the point of convergence between these two values to improve the accuracy of the context determination.

Table 8 gives a statistical overview of the difference between the different models in this investigation. As shown there is a statistical significance between the traditional models (SVM and NB) and the transformer-based models (BERT and RoBERTa). Somewhat interestingly there is statistical significance in the comparison of RoBERTa ($t = 0.45$) and BERT ($t = 0.45$) which shows that in the context of high

Table 8
Paired t-test comparison of model performance on PAN12 and PJ over AST.

Model	SVM	NB	BERT20	BERT45	RoBERTa20
NB	p=0.070	–	–	–	–
BERT20	p=0.009**	p=0.010*	–	–	–
BERT45	p=0.010*	p=0.011*	p=0.883	–	–
RoBERTa20	p=0.009**	p=0.010*	p=0.541	p=0.343	–
RoBERTa45	p=0.010*	p=0.011*	p=0.062	p=0.023*	p=0.052

Table 9
AST and t value optimum for RoBERTa over PAN12.

t Value	AST 0.05	AST 0.10	AST 0.15	AST 0.20
0.45	0.960	0.960	0.958	0.958
0.47	0.961	0.960	0.959	0.959
0.49	0.954	0.956	0.956	0.951

confidence message lines RoBERTa performs significantly better than BERT, however this same conclusion cannot be made when observing lower confidence message lines ($t = 0.2$). This can be explained by the MLA results obtained in Table 2 in which Adult determinations were made with greater accuracy with RoBERTa in comparison to BERT across all t values, and while BERT has some benefit over RoBERTa for child determinations this is not to the same degree as the Adult determinations. It is suggested however that for these adult determinations this could be the result of overfitting as denoted by the IA values for PAN12 as shown in Table 3.

5.2.3. AST & t value regression for FP/FN convergence

Following the findings in this investigation in which RoBERTa is the best-performing model for context determination, it was observed that this was for the greatest AST value and t values that were tested.

Table 9 shows the F1 scores of this analysis and as shown the assumption made to increase the F1 score by increasing the AST value beyond what was previously tested was incorrect, however increasing the t value to 0.47 some minimal increase in F1 score over using a t value of 0.45. This therefore indicates a probable performance ceiling of using the context determination approach over the models used in this investigation, with the use of differing ASTs having limited impact on performance when using these very high confidence t values.

5.3. Discussion

Despite the poor results for the MLA on PAN12 the results from the full context determination of PAN12 suggest that this approach can be used to provide an OG determination with good accuracy. It is proposed that this approach in conjunction with other models can allow for the robust, complex determination of OG. In addition to this a limitation of this investigation is that in the positive OG transcripts the child messages were sent by an adult pretending to be a child, therefore the language used may not be truly representative of how a child would interact in these transcripts. When considering the four models that were used in this investigation it is clear that the transformer-based models were better performing. This was clearest within the MLA results for the same dataset (over PJ sample) with a difference of up to 14.1% in the rate of correct determinations (between NB and RoBERTa with $t = 0.45$). The results for MLA for robustness across the PAN12 dataset were somewhat similar across all models - in terms of correct determination percentage. However, it should be noted that with BERT and RoBERTa, due to the Message Threshold value, there are some omissions within the messages that could alone explain the percentage increase in accuracy as NB and SVM did not incorporate this process. In terms of the impact on the full context determination metric for the PJ sample, we can observe a difference in F1 score of up to 0.14 between NB and RoBERTa ($t = 0.45$), when considering an AST value of 0.001; a difference in F1 score of up to 0.12 between NB and RoBERTa ($t = 0.45$), when considering an AST value of 0.01; and a difference in F1 score of up to 0.11 between NB and RoBERTa ($t = 0.45$), when considering an AST value of 0.05. Despite this reduction in the difference between F1 results across different AST values, the increase in AST value leads to better performance. This is also the case in increasing the t value, most notably when using RoBERTa. It is suggested that increasing this AST value to observe the impact would be useful in finding the best solution, in the context of the PJ sample due to the lack of negative transcripts we are unable to confidently assume that this will increase the overall accuracy. However as the t value does not directly affect this methodological concern in this experiment, this value can be adjusted closer to 0.5 to observe the impact this has on the overall accuracy.

In terms of the context determination for PAN12, in which robustness is being considered, we can see a significant difference in the F1 score between the non-transformer-based models and transformer-based models across the usage of all AST values. This is more prominent when using a smaller AST value. However, it can be observed that an F1 accuracy score of up to 0.96 from using the full PJ dataset in determining OG across PAN12, when using RoBERTa ($t = 0.45$) and an AST of 0.05. This validates the claim that context determination is a robust approach that can be used in the determination of OG. Furthermore as shown by Fig. 8 we can observe the low false positive frequencies in comparison to the false negative frequencies across all models. This suggests that an increase in AST value could lead to a further reduction in false negatives and, despite a probable increase in false positives, should lead to an increase in the overall accuracy. This is in conjunction with using a greater t value. It is suggested that an implementation of this approach would require other insights to provide further evidence/a bespoke approach for the SMP. Due to this proposed combination of insights it is suggested that each model/insight within this would allow for the question 'Does this outcome allow us to say that OG is not occurring?' to be answered. If this is the case then a reduction in the false negative frequency is preferable to ensure that true OG transcripts are dismissed and allow for false positive transcripts to be dismissed using one of the other insights in the proposed model.

In terms of the message threshold, it appears that for BERT this had minimal impact on the overall context determination results and the PAN12 MLA results. However in terms of the PJ sample experiment, there is a slight positive correlation between the t value and the percentage of correct determinations, however, if we were to rely on this metric alone we are likely to not consider the fact that 30.0% of the messages would be omitted (if using BERT, with a t value of 0.45) therefore in terms of a realtime implementation this will require many more messages to make a determination and, in a real-world context, may cause the severe impacts of OG to occur before a determination can be made. This, however, is out of the scope of the full context transcript determination approach. While it is assumed that the omission mechanism discussed in this investigation relates to non-polarised messages, a limitation of this study is the lack of validation of this assumption. Therefore further investigation into these low confidence message lines is required to observe how the approach interprets different phenomena e.g. a message line containing both traditionally 'adult' and 'child' language.

5.3.1. Use case analysis

In the use case of best performance, it is clear from the results obtained that using RoBERTa with an AST of 0.05 and a t value of 0.45 is the best approach. However, when considering the real-world implementation of this approach for SMPs this is dependent on if the results of this approach will be combined with another insight. If this approach were to be implemented alone then it would be preferable to seek for the lowest rate of False Positives to reduce the unnecessary breach of privacy/intervention from human reviewers that would be required as part of incident response. Therefore the results within this investigation

suggest taking an approach with a lower t value and higher AST value would provide the best solution to this (as demonstrated by RoBERTa $t = 0.2$ with an AST of 0.001). However in the case of this insight being used in conjunction with other insights, which is the intention for this approach, then subject to the results of these other insights and further analysis - it is preferable to have a lower False Negative score (which would require the use of the best-performing approach), if an 'evidence to reject' approach is taken, or the greatest accuracy if a Fuzzy logic approach is applied to this complex OG determination approach.

6. Conclusion and future work

This investigation demonstrates that a context determination approach is significantly effective at determining AC interactions as seen within the same (F1 = 0.973) and different (F1 = 0.960) datasets. There is some uncertainty about the generalisability of this approach to non-OG AC contexts however analysis across different use cases can allow for this to be tested. When considering the research questions on which this paper is based:

- ***Can a robust model be formed from taking a full transcript context determination approach?***: While there is some difference between the F1 score obtained from taking the same dataset and a different dataset approach, this difference is minimal (−0.013) and unlikely to impact any real-world implementation of this approach in any meaningful way. This provides hope that this approach will be able to be used across many different social media platform contexts with a good level of effectiveness.
- ***What impact do t values and AST values have on the transcript OG determination accuracy?***: Within this investigation, there was a clear trend that across the best performing model (RoBERTa) an increase in t value led to an overall reduction in the False Negative frequency for the PJ sample experiments and the False Negative and False Positive frequencies across the PAN12 experiments. This suggests that it might be most effective to omit a high proportion of messages sent within this analysis and only rely on the most polarising messages. As shown by the F score regression within Table 9 there seems to be a limit to the polarisation of these messages, as shown by the dip in F1 score when using a t value of 0.49 (which equates to only considering messages between 1–0.99 and 0–0.01), with the best performing t value of 0.47 (which equates to only considering messages between 1–0.97 and 0–0.03). It is expected that in a modern-day implementation, the 'robustness gap' between PJ and PAN12 is smaller than modern-day OG attacks i.e. we expect to see that current attacks have greater linguistic differences to PJ and PAN12 than PAN12 to PJ. In terms of AST values, it can be seen that an increase in this variable leads to an increase in F1 score across both datasets. This appears to be the case due to the greater False Negative frequency in comparison to the False positive frequency that can be seen across the PAN12 experiments, and therefore while decreasing the AST value causes a fall in False Positives there is a smaller initial frequency (at AST = 0.05) compared to the False Negative difference (when comparing AST = 0.001 and AST = 0.05).
- ***What model is recommended to be used in different use cases?***: It is recommended that the transformer-based models are most appropriate for most use cases. The only consideration that should be made to this is computational power/prediction time that is required with these models may make these not fit for purpose. However, it is suggested that while a decent F1 score of 0.871 was obtained with SVM across the PJ sample (when using AST = 0.05), it is clear that this approach lacks robustness as shown by the F1 value obtained across PAN12 for the same experiment (0.748).

6.1. Future work

Further work is required to provide a solution to OG detection, this predominately refers to the continued analysis of psychological literature to provide a robust solution that can detect the complex differences between the variety of attack methods that are used within OG. The generalisability to current Online Grooming attacks is an ongoing limitation with the PJ and PAN12 datasets used in this investigation and throughout literature, future work should consider how to ensure linguistic similarity to modern day communications. For example with the extensive use of 'lol' in these datasets being used by honeypot children is unlikely to be representative of current language patterns (Beck, 2025). This is in addition to making the assumption that the honeypot child in the transcript is an accurate representation of the linguistics used by a real child. Based on the context of the sting operation carried out there were certain behaviours that the honeypot child had to exhibit or not exhibit e.g. suggesting a physical meet up at the sting house and not actively engaging in 'sexting'. In the context of a genuine online grooming attack with a real child these behaviours might be present and therefore these may lead to incorrect Adult-Child classifications if applied to a genuine attack.

Another consideration that should be made in future work is the differentiation of CCSOs and FCSOs in which it is suggested that once establishing an Adult-Child interaction using context determination this could be possible from then determining a 'physical meet-up insight', however a classification task of this type has not yet been attempted in literature. A desirable outcome within a solution would be the ability for real-time detection. The analysis of sealed OG transcripts found that the severe effects (e.g. sending of sexual media) of OG happen at around the 50%–70% point of the transcript (Chiang & Grant, 2019). It should be noted that by taking a real-time approach these additional insights mentioned in this section would unlikely be established, and thus a greater reliance on the context determination method as described in this paper. However despite this future work should observe the computational cost of the Context Determination approach, as with literature considering the latency of detection in a binary predator classification context (An, Ryu, Do, Kim, Ok, & Lee, 2025; Chehbouni, De Cock, Caporossi, Taik, Rabbany, & Farnadi, 2025). It is suggested that by utilising lower confidence t values a smaller sample will be required and thus less determinations on message lines will need to be used, however this assumption needs to be investigated. In addition, due to the increase in prevalence of group chat interactions and other features e.g. destructive messages on Social Media Platforms, the integration of Context Determination within this should be considered. Further work should also analyse the presence and possible mitigations of the 'skew effect' when comparing different dataset sources based on the potential overfitting of the Adult determinations, as seen when comparing the results of MLA between PJ sample and PAN12. The following objectives are recommended as further work in this area:

- Investigating the generalisability of context determination to modern linguistic patterns.
- Identify the real world feasibility of this approach considering computational complexity and latency.
- Utilising the modern day social media platform features in this context determination.

CRediT authorship contribution statement

Jake Street: Conceptualization of this study, Methodology, Investigation, Writing – original draft. **Isibor Kennedy Ihianle:** Supervision, Writing – review & editing. **Funminiyi Olajide:** Supervision, Writing – review & editing. **Ahmad Lotfi:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A github repository with code has been shared in the manuscript.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. software available from tensorflow.org, URL: https://www.tensorflow.org/.

An, J., Ryu, S., Do, H., Kim, Y., Ok, J., & Lee, G. G. (2025). Revisiting early detection of sexual predators via turn-level optimization. arXiv preprint arXiv:2503.06627.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*, 119–123. http://dx.doi.org/10.1145/1461928.1461959.

Beck, S. N. (2025). Im embarrassed 2 say lol: The functions of lol in online groomer-decoy interactions. *Journal of Pragmatics*, *247*, 89–102.

Borj, P., Raja, K., & Bours, P. (2023). Detecting online grooming by simple contrastive chat embeddings. In *Proceedings of the 9th ACM international workshop on security and privacy analytics*. http://dx.doi.org/10.1145/3579987.3586564.

Butler, N., Quigg, Z., & Bellis, M. A. (2020). Cycles of violence in england and wales: the contribution of childhood abuse to risk of violence revictimisation in adulthood. *BMC Medicine*, *18*, 1–13.

Chehbouni, K., De Cock, M., Caporossi, G., Taik, A., Rabbany, R., & Farnadi, G. (2025). Enhancing privacy in the early detection of sexual predators through federated learning and differential privacy. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 27887–27895).

Cheong, Y. G., Jensen, A. K., Gudnadottir, E. R., Bae, B. C., & Togelius, J. (2015). Detecting predatory behavior in game chats. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*, 220–232. http://dx.doi.org/10.1109/tciaig.2015.2424932.

Chiang, E., & Grant, T. (2019). Deceptive identity performance: Offender moves and multiple identities in online child abuse conversations. *Applied Linguistics*, *40*, 675–698. http://dx.doi.org/10.1093/applin/amy007, URL: https://academic.oup.com/applij/article/40/4/675/4952155.

Chiu, M. M., Seigfried-Spellar, K. C., & Ringenberg, T. R. (2018). Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: statistical discourse analysis of self-disclosure and emotion words. *Child Abuse & Neglect*, *81*, 128–138. http://dx.doi.org/10.1016/j.chiabu.2018.04.004.

Department for Education (2023). Teaching online safety in schools. URL: https://www.gov.uk/government/publications/teaching-online-safety-in-schools/teaching-online-safety-in-schools.

Ebrahimi, M., Suen, C. Y., & Ormandjieva, O. (2016). Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, *18*, 33–49. http://dx.doi.org/10.1016/j.diin.2016.07.001.

Edwards, A., Edwards, L., Leatherman, A., Kontostathis, A., Bayzick, J., Mcghee, I., et al. (2009). Comparison of rule-based to human analysis of chat logs comparison of rule-based to human analysis of chat logs.

Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In *Proceedings of the international AAAI conference on web and social media*: *Vol. 3*, (pp. 214–217). http://dx.doi.org/10.1609/icwsm.v3i1.13992.

Hamm, L. (2025). Advancing grooming detection in chat logs: Comparing traditional machine learning and large language models with a focus on predator tone.

HM Government (2021). Tackling child sexual abuse strategy 2021 1 tackling child sexual abuse strategy. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973236/Tackling_Child_Sexual_Abuse_Strategy_2021.pdf.

Home Office (2023). End-to-end encryption and child safety. URL: https://www.gov.uk/government/publications/end-to-end-encryption-and-child-safety/end-to-end-encryption-and-child-safety#:~:text=End%2Dto%2Dend%20encryption%20(.

Inches, G., & Crestani, F. (2012). Overview of the international sexual predator identification competition at pan-2012. URL: https://downloads.webis.de/pan/publications/papers/inches_2012.pdf.

Kloess, J. A., Seymour-Smith, S., Hamilton-Giachritsis, C. E., Long, M. L., Shipley, D., & Beech, A. R. (2015). A qualitative analysis of offenders' modus operandi in sexually exploitative interactions with children online. *Sexual Abuse: A Journal of Research and Treatment*, *29*, 563–591. http://dx.doi.org/10.1177/1079063215612442.

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. In *Applications and theory* (pp. 149–164). John Wiley & Sons, http://dx.doi.org/10.1002/9780470689646.ch8, URL: http://webpages.ursinus.edu/akontostathis/TextMining2009BookChapter.pdf.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692, http://arxiv.org/abs/1907.11692. [arXiv:1907.11692].

Lobe, B., Livingstone, S., Ólafsson, K., Vodeb, H., & Vodeb, K. (2011). Cross-national comparison of risks and safety on the internet: initial analysis from the eu kids online survey of european children report original citation: Lobe. URL: https://eprints.lse.ac.uk/39608/1/Cross-national%20comparison%20of%20risks%20and%20safety%20on%20the%20internet%28lsero%29.pdf.

Michalopoulos, D., & Mavridis, I. (2011). Utilizing document classification for grooming attack recognition.

Milon-Flores, D. F., & Cordeiro, R. L. (2022). How to take advantage of behavioral features for the early detection of grooming in online conversations. *Knowledge-Based Systems*, *240*, Article 108017. http://dx.doi.org/10.1016/j.knosys.2021.108017.

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2021). How old do you think i am? a study of language and age in twitter. In *Proceedings of the international AAAI conference on web and social media*: *Vol. 7*, (pp. 439–448). http://dx.doi.org/10.1609/icwsm.v7i1.14381.

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. http://arxiv.org/abs/2308.14683. [arXiv:2308.14683].

O'Connell, R. (2003). A typology of child cybersexploitation and online grooming practices. URL: https://image.guardian.co.uk/sys-files/Society/documents/2003/07/17/Groomingreport.pdf.

Office of Communications (2023). Children and parents: Media use and attitudes. URL: https://www.ofcom.org.uk/_data/assets/pdf_file/0027/255852/childrens-media-use-and-attitudes-report-2023.pdf.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peersman, C., Daelemans, W., & Vaerenbergh, L. V. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on search and mining user-generated contents*. http://dx.doi.org/10.1145/2065023.2065035.

Rangel, F., & Rosso, P. (2013). Overview of the author profiling task at pan 2013. URL: https://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf.

Street, J. (2025). Online grooming context determination. URL: https://github.com/thejakestreet/Context-Determination.

Street, J., & Olajide, F. (2023). Evaluating a non-platform-specific ocr/nlp system to detect online grooming. In *International conference on cyber warfare and security*: *Vol. 18*, (pp. 504–511). http://dx.doi.org/10.34190/iccws.18.1.967.

Villatoro-Tello, E., Juárez-González, A., Escalante, H., Montes-Y-Gómez, M., & Villaseñor-Pineda, L. (2012). A two-step approach for effective detection of misbehaving users in chats notebook for pan at clef 2012. URL: https://downloads.webis.de/pan/publications/papers/villatorotello_2012.pdf.