Protecting Children from Online Exploitation: Can a Trained Model Detect Harmful Communication Strategies?

Darren Cook*†
Imperial College London
London, United Kingdom
darren.cook@imperial.ac.uk

Miri Zilka*
University of Cambridge
Cambridge, United Kingdom
mz477@cam.ac.uk

Heidi DeSandre
University of Liverpool
Liverpool, United Kingdom

Susan Giles
University of Liverpool
Liverpool, United Kingdom

Simon Maskell
University of Liverpool
Liverpool, United Kingdom

ABSTRACT

The growing popularity of social media raises concerns about children's online safety. Of particular concern are interactions between minors and adults with predatory intentions. Unfortunately, previous research on online sexual grooming has relied on time-intensive manual annotation by domain experts, limiting both the scale and scope of possible interventions. This work explores the possibility of detecting predatory behaviours with accuracy comparable to expert annotators using machine learning (ML). Using a dataset of 6771 chat messages sent by child sex offenders, labelled by two of the authors who are forensic psychology experts, we study how well can deep learning algorithms identify eleven known predatory behaviours. We find that the best-performing ML models are consistent but not on par with expert annotation. We therefore consider a system where an expert annotator validates the ML algorithms outputs. The combination of human decision-making and computer efficiency yields precision—but not recall—comparable to manual annotation, while taking only a fraction of the time needed by a human annotator. Our findings underscore the promise of ML as a tool for assisting researchers in this area, but also highlight the current limitations in reliably detecting online sexual exploitation using ML.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Information systems** → **Top-k retrieval in databases**.

KEYWORDS

Child sexual exploitation, online grooming, chat logs, machine learning, natural language processing

*Authors contributed equally to this research.
†The work was completed while the author was affiliated with the University of Liverpool, UK. The author's affiliation has since changed to Imperial College London, UK.

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604696

ACM Reference Format:
Darren Cook, Miri Zilka, Heidi DeSandre, Susan Giles, and Simon Maskell. 2023. Protecting Children from Online Exploitation: Can a Trained Model Detect Harmful Communication Strategies?. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3600211.3604696

1 INTRODUCTION

Online sexual grooming is an increasing problem in the digital age [25]. In 2021 alone, UK police forces recorded over 5,000 offences relating to sexual communication with a child, representing an increase of over 70% on the three years prior [42]. In the US, 5.4% of adolescents experience online grooming between the ages of 13–17 [23]. The victims of online predators often endure significant harm, with many abusers seeking physical contact offline [48]. Early identification of these predators is therefore crucial.

In prior work on identifying such predators [6, 43, 51, 52], researchers have largely relied on manual annotation of online conversations between predators and their victims. This work is time-consuming and prone to error. While machine learning (ML) has the potential to automate some of this effort, its use for preventing child endangerment online remains underexplored.

We investigate the extent to which ML algorithms can help with detection of online predatory behaviours. One of the involved challenges is that acquiring real-world data featuring minors is inherently difficult due to moral concerns regarding the protection of victims, logistical issues in data collection, and ethical constraints ensuring that data is handled sensitively. Consequently, we utilise a corpus of chat logs from Perverted Justice [21], an online watchdog featuring adult decoys impersonating underage victims.

We focus on identifying 11 communication strategies characteristic of predatory interactions, based on a framework developed, established, and validated by forensic psychologists [18]. While these strategies do not cover all predatory behaviours, they represent many of the actions that law enforcement deems problematic. The 11 behaviours are also subtle and difficult to discern even for experts, leading to frequent disagreement. Therefore, in addition to conventional metrics like precision and recall, we also examined the level of inter-rater agreement, and how it relates to the deviations of machine-generated annotations from the experts.

Section 2 outlines the background and challenges of automatic detection of online predatory behaviours. We then describe our methodology in Section 3.1 and investigate the performance of fully

automated annotation in Section 3.2. The results are unsatisfactory for several behaviours, especially those which appear more rarely in the manually annotated corpus. We address this in Section 4, where we aim to enhance performance via human-computer collaboration. We let the computer extract conversation segments representative of one of the communication strategies before one of the authors, a forensic psychology expert, verifies the resulting predictions. This approach significantly improves the overall precision while maintaining an order of magnitude higher efficiency relative to manual annotation. In Section 5, we address the ethical implications of automated detection of predatory behaviour. Finally, we discuss limitations and summarise our findings in Section 6.

2 BACKGROUND

2.1 Online Child Sexual Exploitation

Existing social science literature on online grooming is extensive, focusing primarily on classifying predatory behaviours in order to aid law enforcement. Researchers have identified two types of sexual predators based on whether they seek to establish physical contact, or wish to engage in fantasy-like discourse [15]. Unsurprisingly, due to the added risk of offline offending, most research has focused on identifying offenders who seek physical contact with their victim [6, 43, 51, 52]. However, the notion that predatory behaviour can be parsed into wholly online or offline offending is oversimplified. For instance, a systematic review of 22 empirical studies found minimal evidence of offenders who solely engage in contact or fantasy-seeking behaviours [8]. Recognising the spectrum of tactics used by offenders is crucial for improving detection methods and tailored interventions.

Current models of online grooming do not account for how the role of the victim impacts the predator's response [14, 15, 33–35, 37]. This lack of understanding in bilateral communication between offender and victim has meant that law enforcement often relies on rudimentary methods such as detecting hyper-sexualised keywords to identify predatory activity. Where the literature has taken a bilateral approach, results have highlighted the importance of understanding the linguistic exchange between offenders and victims. For example, Seymour-Smith and Kloess [47] found that predators would request sexually explicit images in part to trap and control their victims. Once in possession of the images, predators utilised overt persuasion and extortion to overcome victim non-compliance. Such insight demonstrates the utility that can be gleaned from considering the victim's role and the predator's tactics, something that would not be possible by merely scanning for hyper-sexualised keywords, and underscores the necessity for more sophisticated approaches to detecting online predatory behaviour.

Elliot's Self-Regulation model [22] is the first to incorporate victims' behaviour into a model of online predatory grooming. Self-Regulation is a feedback system comprising two phases: a) potentiality, and b) disclosure. *Potentiality* includes mechanisms for rapport-building, incentivising the relationship, disinhibiting the victim, and managing security risks. *Disclosure* primarily concerns whether the predator's behaviour has sufficiently desensitised the victim. Outcomes of this stage can include seeking agreement on a common goal (e.g., arranging offline contact).

As the landscape of online communication continues to evolve, so does the complexity of predatory tactics. While invaluable in providing a foundational understanding, traditional methods have shown limitations in scalability and adaptability to the changing modus operandi of online groomers. This limitation necessitates the exploration of more automated approaches like machine learning.

2.2 Automated Approaches

Offenders use a variety of subtle behaviours to manipulate the conversation flow, such as flattery to build trust [2], or threats and bribes as a coercion tactic [30]. This subtlety can be challenging for automated approaches to detecting predatory behaviours [9]. Prior attempts like [3] relied on dictionary-based approaches, which often result in a large number of both false positives and negatives [7, 31].

Another strain of literature has focused on identifying predators from a mixed corpus of illicit and everyday conversations [20, 26, 27, 29, 41, 45]. While valuable in its own right, this line of research does not offer significant value to law enforcement, as it lacks psychological insight that could justify a preventative intervention. Furthermore, an ML algorithm trained to distinguish between mundane and predatory conversations may overly rely on sexual words [19], while missing more subtle intimacy-seeking, social, and opportunistic behaviours. For example, some predators withhold sexually explicit talk to establish rapport and control [22], or fulfil their fantasy of a a conventional relationship with the victim [24].

Finally, most similar to our work is research on using ML to detect behaviours domain experts regard as problematic. For instance, Gupta et al. [27] used psycholinguistic features to identify six phases of a predatory interaction: friendship forming, relationship forming, exclusivity, risk assessment, sexual activity and conclusion. Similarly, Gunawan et al. [26] used supervised ML to align these phases with specific behaviours such as asking for a picture, talking about friends, discussing hobbies, and building mutual trust. Cano et al. [12] undertook a similar task using a social signal processing approach. Other studies have used a combination of ML and dictionaries to detect qualitative differences in linguistic behaviour between the messages produced by predators and those generated by victims [19], or quantify the level of predatory behaviour from crowd-sourced metrics [45].

While there has been progress in understanding online child sexual exploitation and developing suitable detection methods, a gap remains in addressing the complexity of predatory tactics. Moreover, existing approaches often struggle to detect subtle predatory behaviours and instead rely on detecting sexually-explicit keywords. This study aims to address these gaps by employing advanced machine learning techniques to develop a more nuanced detection model to identify the subtle behaviours predators use throughout the online grooming process.

3 AUTOMATED LABELLING OF COMMUNICATION STRATEGIES

3.1 Method

3.1.1 Dataset. This work uses chat log data between online sexual offenders and adult decoys posing as children and teens. We

Table 1: Behaviour labels used for manual annotation of predator messages, including characteristics of each communication strategy.

Strategy	Code	Characteristics
Communication	COMM	Sustaining the interaction
		Asking questions
		Using linguistic fillers
Rapport	RAPP	Sweet talk
		Show interest
		State shared experiences
Control	CONT	Make demands
		Illusion of victim control
		Ask permission
Challenge	CHAL	Direct confrontation
		Mock insult
		Challenge abilities
Negotiation	NEGO	Arrange to meet
		Offer incentives
Use of Emotion	EMOT	Guilt tripping
		Vilifying third parties
		Playing the victim
Testing Boundaries	TEST	Checks engagement
		Setting boundaries
Sexual Topics	SEX	Stating sexual preferences
		Fantasy talk
		Suggest media production
Mitigation	MITI	Normalising sex
		Downplay age differences
Encouragement	ENCO	Flirting
		Acting as mentor
Risk Management	RISK	Emphasise secrecy
		Acknowledge wrongdoing
		Discuss consequences

compiled twenty-four chat logs from Perverted Justice[1] using an automated web scraping tool built on top of the `beautifulsoup` [46] library in Python. Perverted Justice is a publicly available online repository of two-way instant messaging interactions from sites such as MySpace and Yahoo Instant Messenger. The chats took place between 2003 and 2016. We randomly chose our chat logs from the over 600 available on the Perverted Justice website. On average, chat logs contained 539 messages sent between the two speakers. The interaction would often take place over several days, comprising multiple conversations. The offender always initiated the interaction. The chat logs comprised 12,942 messages in total. Offender messages to the victim accounted for 6,771 (52%) of these.

3.1.2 Data Processing. We extracted chat logs from the Perverted Justice website as plain text files, then inspected and cleaned the data to standardise formatting and remove additional commentary. We also anonymised the text, identifying the speakers only based on their role in the conversation (predator or decoy).

[1]Perverted Justice ceased operations in 2016 but continued to make their data publicly available until March 2023. We originally accessed the data in 2020.

Two of the authors of this work, both possessing a forensic psychology background, used a grounded theoretical approach to label the offender messages. Grounded theory is a flexible methodology designed to extract descriptive (i.e., qualitative) patterns in data [13]. Codes are developed inductively (i.e., data-driven) through an iterative approach to the point of data saturation. The annotators then reviewed and amended prospective codes until they reliably described the interaction. A final coding framework was agreed upon, resulting in eleven communication strategies predators use when responding to their victims. We also included an additional control variable corresponding to a null annotation, i.e., where none of the strategies were found in the respective message. The communication strategies are briefly described in Table 1, and in detail in **??**.

Coding the corpus took four months and over 600 hours to complete. In addition, codes were not mutually exclusive, meaning a predator could display multiple strategies within the same message. This approach and the time-consuming nature of manual coding significantly contributed to the required effort and highlighted the infeasibility of a manual approach for coding large datasets. Based on the time required for this corpus, manually coding the entire corpus held by Perverted Justice would likely require several years of effort.

Due to the time and effort required, it was not feasible to perform repeat coding of our entire corpus. However, inter-rater agreement was sample tested, in addition to collaborative coding exercise during the initial development of the framework. We split the coded predator message corpus into training, testing, and validation regions. 70% was used for training, 20% for testing, and 10% for validation. Data splits were stratified to ensure coverage in each region mirrored that observed in the full corpus. Table 2 reports the distribution of messages per region.

Table 2: Split of predator-to-victim messages in our dataset into training, testing, and validation regions using a 70-20-10 ratio. Splits were stratified to ensure distribution of labels in each region matched the full corpus.

Region	Messages	Data Split
Train	4712	70%
Test	1355	20%
Validation	704	10%

3.1.3 Models. We used a natural language inference (NLI) approach to predict how messages relate to communication strategies. NLI is an NLP technique that focuses on comparing two statements of the text. Specifically, determining whether a given statement (the hypothesis) is inferred or contradicted by another statement (the premise) [5]. If the hypothesis can be inferred from the premise, the relationship is one of entailment. On the other hand, contradiction or neutral outcomes occur when we cannot infer a relationship between the two statements.

In this work, we use each predator message as a hypothesis and form one premise from each communication strategy. For example, "This message is an example of control" would be used for the control strategy [32]. We used each message/label sentence pair during training as input to a deep learning model. We used a version of

`RoBERTa-large` [39] hosted on Huggingface, with an implementation built in `Pytorch` [44]. In addition to pretraining, this model has been fine-tuned for NLI tasks using the Multi-Genre Natural Language Inference corpus [50]. We performed further fine-tuning using our training and validation sets. Model parameters are identical to [49]. We trained our models for 10 epochs with batch size 32 and a learning rate of 10^{-5}.

Model predictions for each message in the test set were binarized by finding an optimal threshold, i.e., one that maximizes correlation with the actual labels, as in [32]. This means we set a different threshold per label, allowing us to achieve better results compared to a pre-determined value (such as a universal 50% cut-off).

3.1.4 Comparing zero-shot and few-shot learning. As the time required to manually label our corpus is a bottleneck that hinders the mobility of our approach to larger datasets, we were interested in how prediction performance suffered when we used a reduced training sample. In addition to training on the entire training set, we experimented with few-shot and zero-shot conditions. In the zero-shot condition, we made predictions on the test set with no additional training. In few-shot settings, we experimented with different amounts of positive training examples between 5 and 150. As before, we used a stratified approach when sampling the positive classes to ensure that the class distribution in the few-shot settings matched the actual distribution of the whole training set.

3.1.5 Expanding the contextual window. We also tested whether the surrounding messages increased the contextual understanding of the model. To examine this, we expanded the message window to include multiple prior messages sent by both speakers, and concatenated them into a single input. In addition to the single message input, we experimented with five-message windows. The five-message window combines each predator message with the two preceding victim and predator messages.

3.2 Experiments and results

3.2.1 Coverage Statistics. We report coverage statistics for each communication strategy in Table 3. We calculate coverage as the proportion of messages with a positive class label. Each of the eleven behaviour codes is highly imbalanced. Except for 'communication' (coverage = 73%), positive labels form the minority class. Inspection of messages that were labelled with the 'communication' label revealed that predators were engaging in considerable amounts of both information-sharing and information-gathering. This was particularly prominent at the beginning of conversations, and characterised by a series of targeted, and directive questioning: "asl?"[2], "are you there alone?", "do you want to give me your number?". There were also a considerable number of attempts to use humour-related acronyms (i.e., "lol", "LMAO", "hehe") that explained the high coverage of 'communication' throughout the corpus.

By contrast, mitigation was the rarest label and appeared in only 3% of predator messages. Aside from communication, the average coverage of the remaining labels was 14.3%, suggesting behaviours appeared rarely. Equally, however, the majority of predator messages (92%) were labelled with at least one behaviour. Threshold values were similarly broad—thresholds for communication, control, challenge and encouragement were all within 0.005, while, rapport, use of sex, and risk management all generated a threshold ≥ .85.

[2]The phrase 'asl' is text-speak for "age, sex, location?"

Table 3: Coverage statistics of each communication strategy over all offender messages. Coverage represents the percentage of messages in the dataset that use the corresponding strategy. Train and Test columns indicate the number of manually labelled positive class instances in the train and test regions, respectively. The threshold column shows the optimized threshold based on the largest cross-validated Matthews correlation coefficient (MCC) between predicted and actual labels.

Strategy	Coverage (%)	Train	Test	Threshold
Communication	73	3445	991	.002
Rapport	15	718	206	.98
Control	21	979	282	.004
Challenge	5	211	60	.005
Negotiation	21	986	283	.75
Use of emotion	16	773	222	.71
Testing boundaries	31	1470	423	.78
Use of sex	18	861	248	.98
Mitigation	3	144	41	.7
Encouragement	8	378	109	.004
Risk management	5	217	62	.88

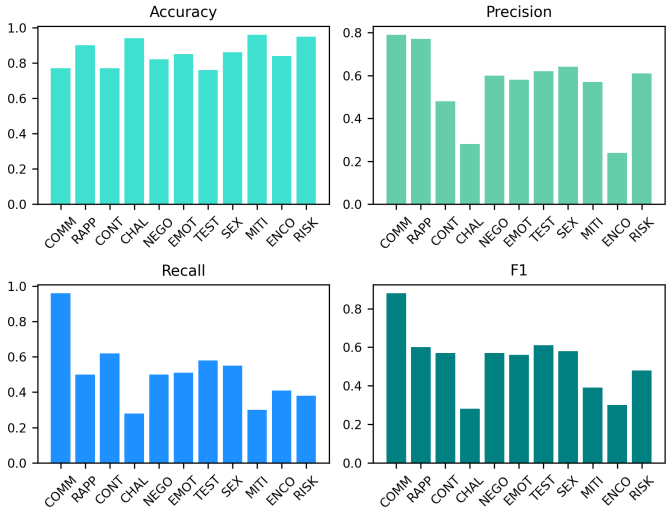

Figure 1: Performance metrics for NLI models trained on all available data for each communication strategy. The subplot shows accuracy, precision, recall, and F1 scores for offender messages within the test set.

3.2.2 Classification of predatory behaviour when trained on all available data. Figure 1 reports the performance of each label when trained with all available training data. Seven of the eleven behaviours also obtain an $F1$ score above 50%, with the best-performing behaviour being 'communication' ($F1$ = .87), followed by 'testing

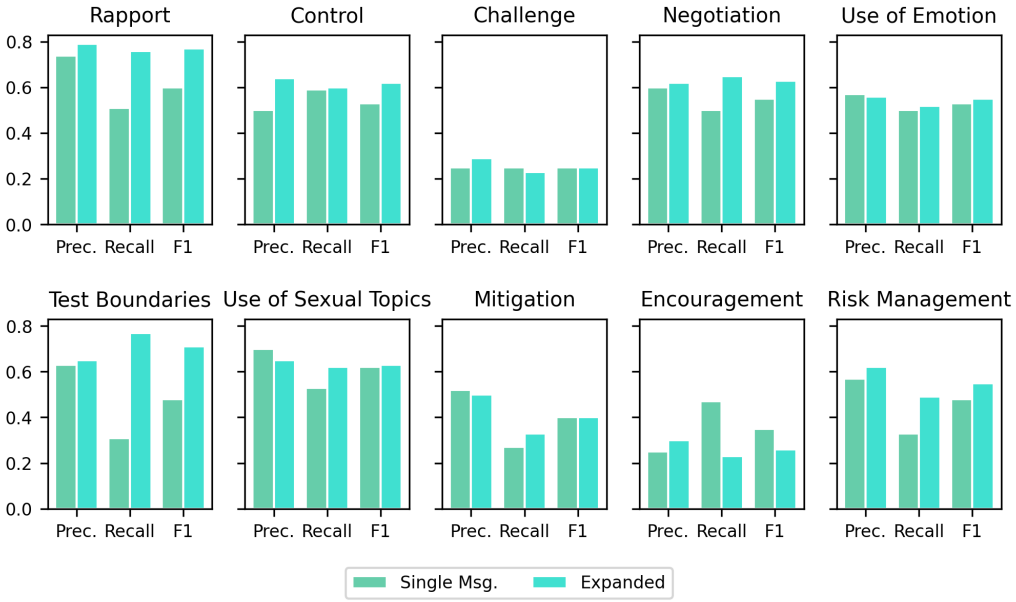

Figure 2: Comparison of precision, recall, and F1 scores for NLI models trained on single and expanded message inputs for each communication strategy. Each subplot displays the evaluation metrics for a specific communication strategy. The x-axis represents the performance metric, while the y-axis represents the score for the metric. The green bars represent the scores achieved by the model trained on a single message input, while the turquoise bars represent the scores achieved by the model trained on an expanded window of 5 messages.

boundaries' ($F1 = .62$), 'rapport' ($F1 = .61$), and 'use of sexual topics' ($F1 = .61$). Performance was poorest for 'challenge' ($F1 = .28$), followed by 'encouragement' ($F1 = .32$), and 'mitigation' ($F1 = .4$). Precision was an issue for the majority of labels, indicating a number of false positives and suggesting that the model had learned some rules that were contributing to a high false-positive rate. To better understand performance, a brief qualitative exploration was performed on a random sample of ≈ 10% of the test set.

The rapport model correctly recognised complements and sweet talk as positive examples, but missed more everyday examples of rapport building such as social greetings, (e.g., "hi, how are you? asl?"). It also routinely failed to identify general conversational patter as evidence of rapport (e.g., "how was your spring break?").

Some aspects of control appeared to take place over longer ranges than single messages. For example, persistently asking the same question was often misclassified, as each message was considered an independent event.

In trying to predict encouragement, which was amongst the worst performing labels, the model appeared to overfit on short verbal nods (i.e., "kool" and "sure"). This appeared regularly in predator speech, but was not always labelled as encouragement by our annotators. Over-reliance on these phrases seems to substantially increase the false positive rate. Risk management appeared to perform better than other rare behaviours. Examination of the positive classifications indicated that this was largely a consequence of recognising attempts to establish the presence of a parent, (e.g.,"is ur dad gona be home tomoro?" and "when are they getting home?").

3.2.3 Comparing classification accuracy with an expanded message window. Figure 2 reports a per behaviour comparison of precision, recall, and F1 between single and multi-message input. Due to the high performance and coverage of the communication strategy, we dropped this label from the remainder of our analysis. The general performance increase was marginal for model precision. However, rapport, control, risk management, and testing boundaries all increased when we included the additional context. However, 'use of sexual topics' decreased precision by 5% (from 70% to 65%) when we used a multi-message window.

An expanded message window markedly increased the recall of several behaviours, including rapport (increased from 52% to 81%), negotiation (from 50% to 67%), testing boundaries (from 31% to 76%), and risk management (from 37% to 48%). This suggests that the added context from the previous messages decreased the false negatives for these behaviours.

3.2.4 Comparing classification accuracy in few-shot and zero-shot conditions. Figure 3 reports the change in the $F1$ as the number of positive training examples increases. At zero-shot, all categories had an F1 score below 50% and half were below 15%. The subjectivity of the behaviours is a possible cause of lacking performance. As noted by [28], concepts such as "rapport" are tough to define, even for humans. It is therefore not surprising that a machine fails at this task without any positive examples for training.

As demonstrated in Figure 3, however, most behaviours notably improved with a small amount of positive training examples. On

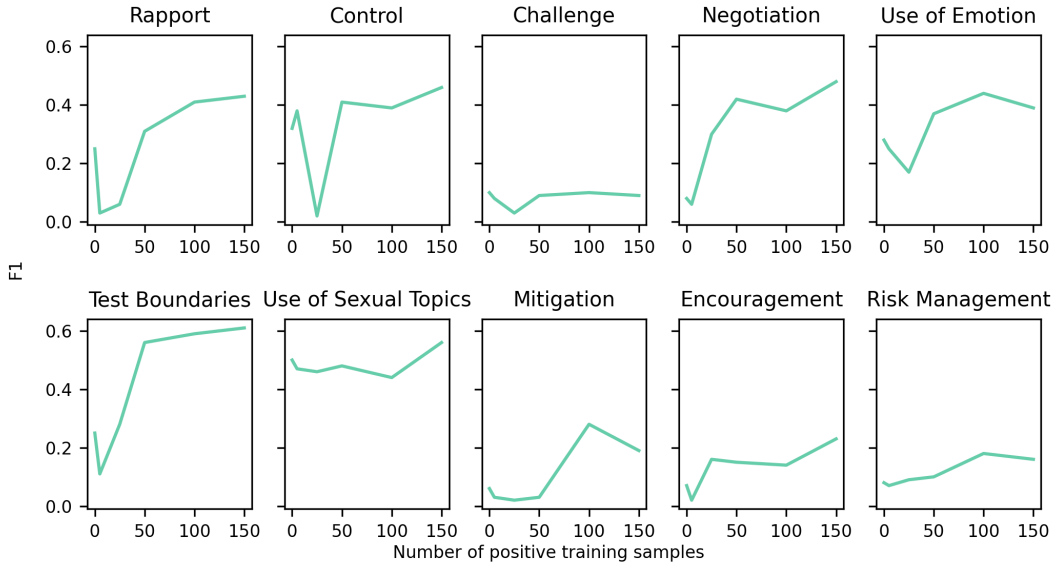

Figure 3: Change in F1 score as the size of the training set increases for each communication strategy. The x-axis represents the number of manually labelled positive instances in the training set, ranging from 0 (zero-shot) to 150. The y-axis represents the F1 score. Each subplot shows the change in F1 score as a line per communication strategy.

average, results indicate that the model attained considerable improvement by training on 50–100 positive examples.

3.2.5 Comparing pairwise agreement between machine and expert annotation. We performed a validation study to explore differences in annotations generated by our forensic psychology experts with those generated automatically by our models. In total, the first author of this work validated the classifications of 645 messages. This step generated a third set of annotations and was deemed a more efficient alternative to re-labelling the corpus from scratch. Cohen's K [16], a standard metric, was used for measuring pairwise agreement between annotators, where larger values of K indicate more agreement between raters. An acceptable level of agreement is subjectively defined. However, social scientists often use the interpretation provided by [38]. In our case, we take the agreement between the two human annotators as the level of 'acceptable' agreement.

Figure 4 reports pairwise agreement scores for each behaviour. Across all behaviours, and for each combination of raters, including the automated one, values of K ranged between .46 and .95, indicating a minimum of 'moderate' agreement on any pairwise combination. Comparing H1 (initial annotations) with H2 (validations performed by the first author), several of the behaviours received a K value above .8, indicating near-perfect levels of agreement. Comparing these agreement scores with those generated by human-machine comparisons (H1 & AI or H2 & AI), values of K are significantly and consistently lower. This finding suggests that our best-performing model was unable to achieve an agreement comparable to an additional human rater. For example, the average agreement between H1 and H2 for 'risk management', 'mitigation', 'negotiation', and 'challenge' was .91 – comfortably within the 'near perfect' range. Conversely, the average agreement between AI and H1 or H2 on the same behaviours was .58. We note that despite being trained on data only from H1, the model did not systemically agree with H1 more than with H2.

4 HUMAN-MACHINE COLLABORATION FOR DETECTION OF PREDATORY COMMUNICATION STRATEGIES

The results presented in Section 3 indicate that whilst a ML solution offer a significant improvement in performance when sufficient training data is available, model precision remains an issue for most behaviours. Over-prediction can result in lost time in high-stakes settings where precision is essential. While in the case of online grooming, it is arguably more tolerable to misidentify non-predatory behaviour as predatory (i.e., lower precision) than to identify predatory behaviour as non-predatory (i.e., lower recall), law enforcement will sacrifice considerable resource unnecessarily if detection of predatory behaviour is consistently poor. It is, therefore, vital that automated systems address this.

This section examines the potential of resolving this issue via a human-in-the-loop approach. While human experts can identify contextual nuances and subtle behaviours that machines may miss, the manual effort required for such annotation is time-consuming, and thus not scalable to large chat datasets. On the other hand, machines need a fraction of the time for processing but, as we have seen, do not achieve the required accuracy. Following [10], we therefore use a weak supervision approach, where the machine is tasked with identifying relevant segments of the chat log, which are then verified by a human expert.

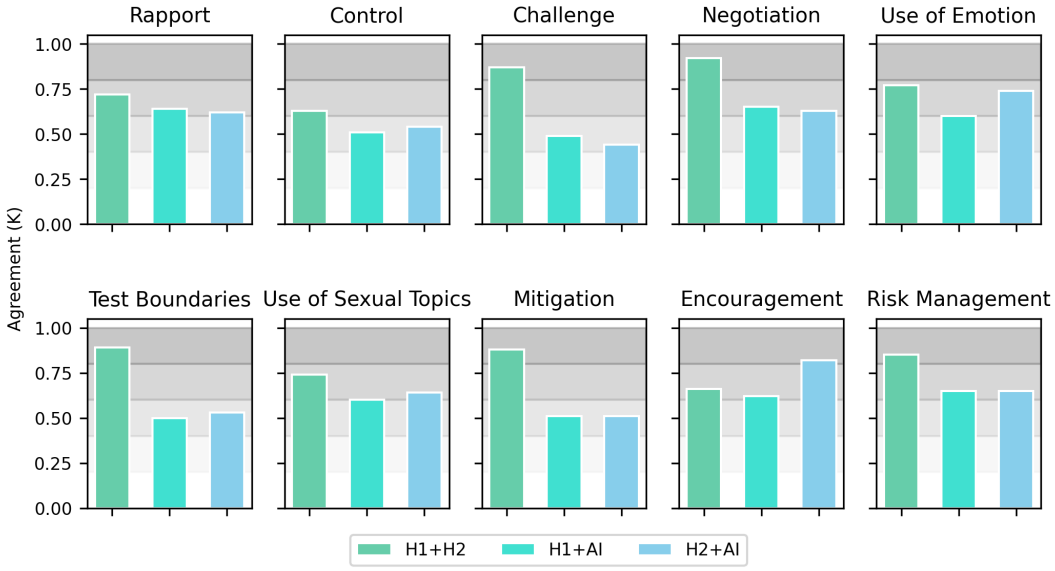

Figure 4: Pairwise agreement between the original human annotations (H1), human verified annotations (H2), and machine-generated annotations (AI) for each communication strategy. The x-axis shows each pairwise combination of raters, and the y-axis displays the Cohen's kappa score, a measure of inter-rater agreement. Shaded areas indicate the level of agreement, ranging from almost perfect (dark grey) to slight agreement (white), according to the interpretation in [38]. Each subplot corresponds to a single communication strategy. The results demonstrate the extent of agreement between different raters and provide insight into the quality of machine-generated annotations compared to human-verified annotations

4.1 Method

4.1.1 Dataset. We used the same twenty-four chat logs used in Section 3, and trained our models with the same data split. In addition, we labelled a further fifteen chat logs from Perverted Justice to increase the size of the test set. In total, we annotated 12,426 messages sent by an offender.

Instead of predicting behaviours at the message level, we grouped messages occurring within a set period into conversations, defined as a continuous sequence of messages where the gap between two messages did not exceed one hour. This step generated sixty-seven conversations, with an average of 185.46 (SD = 188.77) offender messages per conversation. For each conversation and each behaviour label, we extract the conversation segment that best represents each label. This means that the resulting labels indicate which communication strategies were present in each conversation at least once. Note that we omitted the communication category from this analysis as it is likely to be present in all conversations.

4.1.2 Task. We performed our analysis on each conversation within the expanded test set. For each conversation, we used an ensemble of labelling functions–automated methods to annotate data—to extract the segment of text that best represented each behaviour label. Extracted segments were then ranked according to their confidence level, with the top-k segments passed to a human verifier (the first author) to either accept or reject.

4.1.3 Schema. We constructed five labelling functions to extract the text segments. These were: (i) NLI sequence classifier fine-tuned on our training set, (ii) Zero-Shot Q&A classifier, (iii) Zero-Shot Q&A classifier with cosine similarity, (iv) Sentence embeddings with cosine similarity, (v) keyword detection. We provide a complete overview of each of these labelling functions in Appendix **??**.

4.2 Experiments and results

Figure 5 reports performance in precision and recall when $k = 1$ compared to $k = 3$, i.e., when the human validator saw only the model's best guess ($k = 1$) or the top three ($k = 3$). Precision performance was generally very high for both $k = 1$ and $k = 3$, with two labels (Control and Negotiation) obtaining perfect precision when compared to manual annotation. The average precision score across all behaviours was similar, with both conditions performing ≈ 0.94. Given the subjective nature of the labels, imperfect precision mostly corresponds to disagreement between annotators. The lowest performing behaviour was 'Challenge', which dropped by 13% (from 0.8 to 0.73) between the two conditions. This drop in performance is likely due to the $k = 3$ model incorrectly providing more information to the user to verify, thus increasing the likelihood of a false positive. Overall, our findings suggest that a human-in-the-loop approach can consistently extract relevant text segments for the user to review.

However, for most categories, the collaborative set up did not improve recall compared to fully automated methods. In the present context, low recall (an excess of false negatives) can be explained either as a consequence of inter-annotator disagreement (i.e., the model provided excerpts that the verifier rejected, in disagreement with the original annotation), or an inability of the model to identify salient information for a given category (i.e., the model fails to

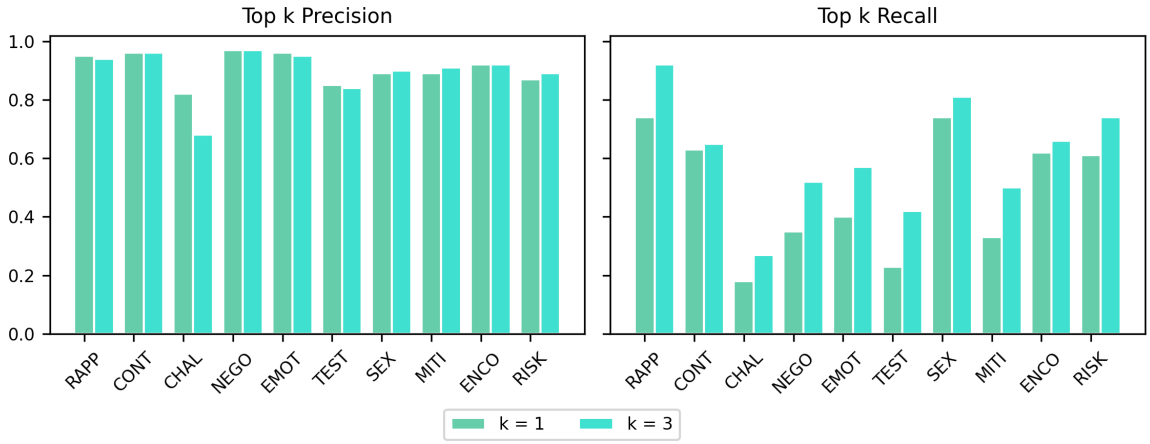

Figure 5: Comparison of precision and recall scores for each communication strategy using top k extraction. The AI-generated evidence was manually verified by humans. High precision scores indicate that the AI-selected evidence aligned with human interpretation of each communication strategy. Lower recall scores suggest that the AI may not have identified all relevant evidence for human review. The green and turquoise bars represent top k extraction with k=1 and k=3, respectively

return anything for the user to verify). Supporting the latter explanation, providing more information for the user to review by increasing the value of k did improve recall for all categories. The average improvement in recall was 11%, from 0.52 ($k = 1$) to 0.63 ($k = 3$).

With respect to the inter-rater disagreement artificially lowering recall, removing such effects would typically require manually re-annotating the entire corpus. However, the high time intensity of manual annotation meant this was not possible in the present context. As an efficient alternative, the first author manually inspected points of disagreement between the original annotations and the output from Section 4.1.2. The first author then re-coded original annotations, and performance metrics were recalculated. Figure 6 reports F1 scores with these amended annotations.

5 ETHICAL AND SOCIETAL CONSIDERATIONS

Developing frameworks to support the automatic detection of online grooming raises critical ethical considerations. For example, data acquisition regulations may hinder accessing the large volumes of data required to train a machine learning algorithm suitably [4]. Other cyber-security issues, such as proper data storage and the potential for hacking, also mean that law enforcement is often reluctant to release actual investigative material, such as chatlogs, for academic purposes [36]. Other privacy issues behind using actual investigative data include difficulties obtaining informed consent for bulk data collection [40].

This work utilises a large corpus of online predatory chat logs archived by a child-safety watchdog organisation. Both the creation and use of this data are controversial. For example, the Perverted Justice model has been criticised for encouraging cyber-vigilantism [54]. Moreover, the fact that offline meetings were routinely televised as part of NBC's To Catch a Predator series has resulted in claims of unnecessary humiliation towards individuals who, at that point, had been neither charged nor convicted of a crime [1]. Additionally, debate exists around whether the persistence of some volunteers constituted legal or moral entrapment [11, 21].

While using the Perverted Justice corpus raises ethical questions, it is important to consider the context in which this data was collected. All predators featured in the chat logs were later convicted of a crime (according to the administrators of the Perverted Justice website, the undercover volunteers' activities resulted in the criminal conviction of over 600 predators between 2003 and 2016). Decoys did not initiate contact with the offender or introduce sexual content, and the conversations did not feature children but an adult playing the role of an underage victim. Notwithstanding these ethical challenges, the difficulties associated with accessing chat logs with real victims have meant that the Perverted Justice archives have become a viable and effective alternative.

Another ethical consideration is the risk of perpetuating harm towards children who have experienced sexual abuse, if researchers mishandle the data used to train an algorithm. There is a need to consider the potential impact of using sensitive information, and to ensure that the rights and dignity of children are respected. Additionally, use of automated detection systems may have unintended consequences, such as false positives or misidentification, which can lead to unjust accusations and damage to innocent individuals' reputations [53]. However, in a deployed system, false negatives are more severe, as they may prevent law enforcement from saving a child from harm.

Due to the high-risk nature of this application, and the level of performance our system achieves, it is clear the technology is not ready to reliably assist in detecting online grooming behaviour in the real world. However, it shows promise in helping researchers working on this crucial domain streamline and speed up their annotation process. Annotating large volumes of text data containing potentially disturbing content can be emotionally challenging. While research that explicitly explores annotator well-being is scarce,

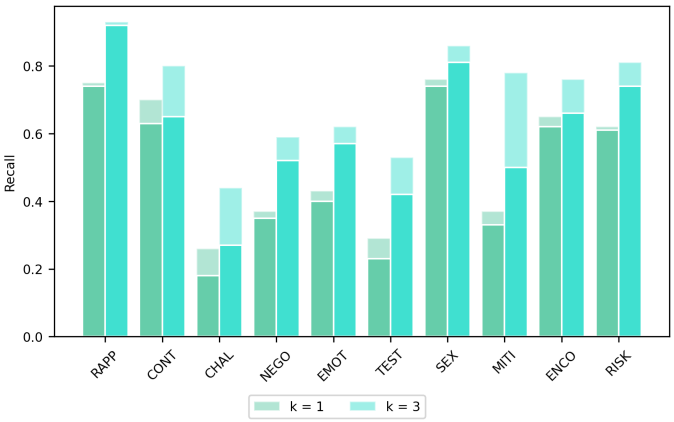

Figure 6: Comparison of adjusted recall scores for each communication strategy after resolving disagreement between original annotations and human verified annotations. Original annotations and human verified annotations were re-annotated by the first author to account for potential discrepancies between the two sources. Bars represent the recall score, where an increase indicates improved recall after adjusting annotations. The analysis aims to investigate whether low recall scores are due to differences in annotators or an inability of the AI to generate relevant evidence. Top 3 extraction, where k = 1 (green) and k = 3 (turquoise) respectively, was used to automatically extract segments of messages as evidence of each communication strategy for human verification.

a related area that has received attention is the role of content moderators on social media. Research has shown that prolonged exposure to harmful material can cause psychological distress, such as post-traumatic stress disorder [17]. The development of automated systems could be helpful in proactively protecting the mental well being of those on the front-lines of data annotation.

6 CONCLUSIONS

Manually labelling the 24 chat logs used in this work took over 600 hours. Given that the full Perverted-Justice corpus contains 850 chat logs, it would be infeasible to label the entire corpus without the help of automated methods. We find that an ML based approach shows potential when applied to the detection of online predatory behaviour. However, even with training, the agreement between the model and a human annotator is not comparable to the agreement between two human annotators.

Adding a human validation step to the annotation process improves precision significantly for the cost of a small-time investment compared to human annotation. However, recall remains an issue even in the collaborative setting. Issues in predicting the correct behaviours seem to stem from the rarity of certain behaviours, but also due to their nuanced nature. These conclusions may be transferable to other contexts and annotation schemes involving highly-subjective class labels. Performing post-validation on the automatic classifications allowed us to gain qualitative insight into the model's performance, which may be used to design better prompts and improve performance further. Overall, our results are an encouraging step towards building tools that may assist researchers within this domain, even if the current capabilities are insufficient to build a sufficiently reliable automated model for detection of online sexual exploitation in the real world.

ACKNOWLEDGMENTS

Given the nature of our topic, ethical approval was sought for the initial data collection and manual labelling of the Perverted Justice corpus. Ethical approval was granted by the University of Liverpool's Institute of Population Health Research Ethics Committee (REF: 9972). This project was supported by the Leverhulme Centre for the Future of Intelligence project RC-2015-067.

We are grateful to Jiri Hron of the University of Cambridge for providing critical insights and invaluable feedback on the final draft of this paper.

REFERENCES

[1] Amy Adler. 2011. To catch a predator. *Colum. J. Gender & L.* 21 (2011), 130.
[2] Connie S Barber and Silvia Cristina Bettez. 2021. Exposing patterns of adult solicitor behaviour: towards a theory of control within the cybersexual abuse of youth. *European Journal of Information Systems* 30, 6 (2021), 591–622. https://doi.org/10.1080/0960085x.2020.1816146
[3] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. 2014. Exploring high-level features for detecting cyberpedophilia. *Computer speech & language* 28, 1 (2014), 108–120. https://doi.org/10.1016/j.csl.2013.04.007
[4] Parisa Rezaee Borj, Kiran Raja, and Patrick Bours. 2022. Online grooming detection: A comprehensive survey of child exploitation in chat logs. *Knowledge-Based Systems* (2022), 110039.
[5] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015). https://doi.org/10.18653/v1/d15-1075
[6] Peter Briggs, Walter T Simon, and Stacy Simonsen. 2011. An exploratory study of internet-initiated sexual offenses and the chat room sex offender: Has the internet enabled a new typology of sex offender? *Sexual Abuse* 23, 1 (2011), 72–91. https://doi.org/10.1177/1079063210384275
[7] Laura Jayne Broome, Cristina Izura, and Jason Davies. 2020. A psycho-linguistic profile of online grooming conversations: A comparative study of prison and police staff considerations. *Child Abuse & Neglect* 109 (2020), 104647. https://doi.org/10.1016/j.chiabu.2020.104647
[8] Laura Jayne Broome, Cristina Izura, and Nuria Lorenzo-Dus. 2018. A systematic review of fantasy driven vs. contact driven internet-initiated sexual offences: Discrete or overlapping typologies? *Child abuse & neglect* 79 (2018), 434–444.
[9] Louisa Buckingham and Nusiebah Alali. 2020. Extreme parallels: a corpus-driven analysis of ISIS and far-right discourse. *Kōtuitui: New Zealand Journal of Social Sciences Online* 15, 2 (2020), 310–331. https://doi.org/10.1080/1177083x.2019.1698623
[10] Bradley Butcher, Miri Zilka, Darren Cook, Jiri Hron, and Adrian Weller. 2023. Optimising Human-Machine Collaboration for Efficient High-Precision Information Extraction from Text Documents. *arXiv preprint arXiv:2302.09324* (2023).
[11] Ronald R Butters, Tyler Kendall, and Phillip Carter. 2014. Internet Traps and the Creation of Linguistic Crimes: Perverted Justice as Broadcast Entertainment. *Internet Traps and the Creation of Linguistic Crimes: Perverted Justice as Broadcast Entertainment* (2014), 223–240.
[12] Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani. 2014. Detecting child grooming behaviour patterns on social media. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings 6*. Springer, 412–427.
[13] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.
[14] Emily Chiang and Tim Grant. 2019. Deceptive identity performance: Offender moves and multiple identities in online child abuse conversations. *Applied Linguistics* 40, 4 (2019), 675–698.
[15] Ming Ming Chiu, Kathryn C Seigfried-Spellar, and Tatiana R Ringenberg. 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 (2018), 128–138.
[16] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/

001316446002000104
[17] Cambridge Consultants. 2019. Use of AI in online content moderation. https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf
[18] Heidi DeSandre. 2021. *Bilateral communication between online child sex offenders and decoy children: A qualitative approach.* Master's thesis. University of Liverpool.
[19] Michelle Drouin, Ryan L Boyd, Jeffrey T Hancock, and Audrey James. 2017. Linguistic analysis of chat transcripts from child predator undercover sex stings. *The Journal of Forensic Psychiatry & Psychology* 28, 4 (2017), 437–457. https://doi.org/10.1080/14789949.2017.1291707
[20] Mohammadreza Ebrahimi, Ching Y. Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep Convolutional Neural Networks. *Digital Investigation* 18 (2016), 33–49. https://doi.org/10.1016/j.diin.2016.07.001
[21] Vincent Egan, James Hoskinson, and David Shewan. 2011. Perverted justice: A content analysis of the language used by offenders detected attempting to solicit children for sex. *Antisocial behavior: Causes, correlations and treatments* 20, 3 (2011), 273297.
[22] Ian A Elliott. 2017. A self-regulation model of sexual grooming. *Trauma, Violence, & Abuse* 18, 1 (2017), 83–97. https://doi.org/10.1177/1524838015591573
[23] David Finkelhor, Heather Turner, and Deirdre Colburn. 2022. Prevalence of online sexual offenses against children in the US. *JAMA network open* 5, 10 (2022), e2234471–e2234471.
[24] Petter Gottschalk, Christopher Hamerton, Petter Gottschalk, and Christopher Hamerton. 2022. Online Grooming. *White-Collar Crime Online: Deviance, Organizational Behaviour and Risk* (2022), 219–243.
[25] Emily A Greene-Colozzi, Georgia M Winters, Brandy Blasko, and Elizabeth L Jeglic. 2020. Experiences and perceptions of online sexual solicitation and grooming of minors: A retrospective report. *Journal of child sexual abuse* 29, 7 (2020), 836–854.
[26] Fergyanto E Gunawan, Livia Ashianti, and Nobumasa Sekishita. 2018. A simple classifier for detecting online child grooming conversation. *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 16, 3 (2018), 1239–1248.
[27] Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. 2012. Characterizing pedophile conversations on the internet using online grooming. *arXiv preprint arXiv:1208.4324* (2012). https://doi.org/10.48550/arXiv.1208.4324
[28] Frederick J Heide. 2013. "Easy to sense but hard to define": Charismatic nonverbal communication and the psychotherapist. *Journal of Psychotherapy Integration* 23, 3 (2013), 305.
[29] Giacomo Inches and Fabio Crestani. 2012. Overview of the International Sexual Predator Identification Competition at PAN-2012.. In *CLEF (Online working notes/labs/workshop)*, Vol. 30.
[30] Malin Joleby, Carolina Lunde, Sara Landström, and Linda S Jonsson. 2021. Offender strategies for engaging children in online sexual activity. *Child Abuse & Neglect* 120 (2021), 105214. https://doi.org/10.1016/j.chiabu.2021.105214
[31] Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive Content Detection in Online User-Generated Data: A survey. *Procedia Computer Science* 189 (2021), 274–281. https://doi.org/10.1016/j.procs.2021.05.098
[32] Christoph Kecht, Andreas Egger, Wolfgang Kratsch, and Maximilian Röglinger. 2021. Event Log Construction from Customer Service Conversations Using Natural Language Inference. In *2021 3rd International Conference on Process Mining (ICPM)*. IEEE, 144–151.
[33] Juliane A Kloess, Catherine E Hamilton-Giachritsis, and Anthony R Beech. 2017. A descriptive account of victims' behaviour and responses in sexually exploitative interactions with offenders. *Psychology, Crime & Law* 23, 7 (2017), 621–632.
[34] Juliane A Kloess, Catherine E Hamilton-Giachritsis, and Anthony R Beech. 2019. Offense processes of online sexual grooming and abuse of children via internet communication platforms. *Sexual Abuse* 31, 1 (2019), 73–96.
[35] Juliane A Kloess, Sarah Seymour-Smith, Catherine E Hamilton-Giachritsis, Matthew L Long, David Shipley, and Anthony R Beech. 2017. A qualitative analysis of offenders' modus operandi in sexually exploitative interactions with children online. *Sexual Abuse* 29, 6 (2017), 563–591.
[36] April Kontostathis, Lynne Edwards, and Amanda Leatherman. 2010. Text mining and cybercrime. *Text mining: Applications and theory* (2010), 149–164.
[37] Kamil Kopecký. 2017. Online blackmail of Czech children focused on so-called "sextortion"(analysis of culprit and victim behaviors). *Telematics and Informatics* 34, 1 (2017), 11–19.
[38] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019). https://doi.org/10.48550/arXiv.1907.11692
[40] Danuta Mendelson. 2017. Legal protections for personal health information in the age of Big Data–a proposal for regulatory framework. *Ethics, Medicine and Public Health* 3, 1 (2017), 37–55.
[41] Md Waliur Rahman Miah, John Yearwood, and Sid Kulkarni. 2011. Detection of child exploiting chats from a mixed chat dataset as a text classification task. In *Proceedings of the Australasian Language Technology Association Workshop 2011*. 157–165.
[42] NSPCC. 2021. Record high number of recorded grooming crimes lead to calls for stronger online safety legislation. https://www.nspcc.org.uk/about-us/news-opinion/2021/online-grooming-record-high
[43] Rachel O'Connell. 2003. A typology of child cybersexploitation and online grooming practices. http://image.guardian.co.uk/sys-files/Society/documents/2003/07/17/Groomingreport.pdf
[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[45] Nick Pendar. 2007. Toward spotting the pedophile telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*. IEEE, 235–241. https://doi.org/10.1109/icsc.2007.32
[46] Leonard Richardson. 2007. Beautiful soup documentation. *April* (2007).
[47] Sarah Seymour-Smith and Juliane A Kloess. 2021. A discursive analysis of compliance, resistance and escalation to threats in sexually exploitative interactions between offenders and male children. *British journal of social psychology* 60, 3 (2021), 988–1011.
[48] Joy Shelton, Jennifer Eakin, Tia Hoffer, Yvonne Muirhead, and Jessica Owens. 2016. Online child sexual exploitation: An investigative analysis of offender characteristics and offending behavior. *Aggression and violent behavior* 30 (2016), 15–23. https://doi.org/10.1016/j.avb.2016.07.002
[49] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690* (2021). https://doi.org/10.48550/arXiv.2104.14690
[50] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).
[51] Rebecca Williams, Ian A Elliott, and Anthony R Beech. 2013. Identifying sexual grooming themes used by internet sex offenders. *Deviant Behavior* 34, 2 (2013), 135–152. https://doi.org/10.1080/01639625.2012.707550
[52] Georgia M Winters, Leah E Kaylor, and Elizabeth L Jeglic. 2017. Sexual offenders contacting children online: an examination of transcripts of sexual grooming. *Journal of sexual aggression* 23, 1 (2017), 62–76. https://doi.org/10.1080/13552600.2016.1271146
[53] Aleš Završnik. 2021. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology* 18, 5 (2021), 623–642.
[54] Andreas Zingerle. 2015. Scambaiters, human flesh search engine, perverted justice, and internet haganah: Villains, avengers, or saviors on the internet. In *ISEA Conference*.