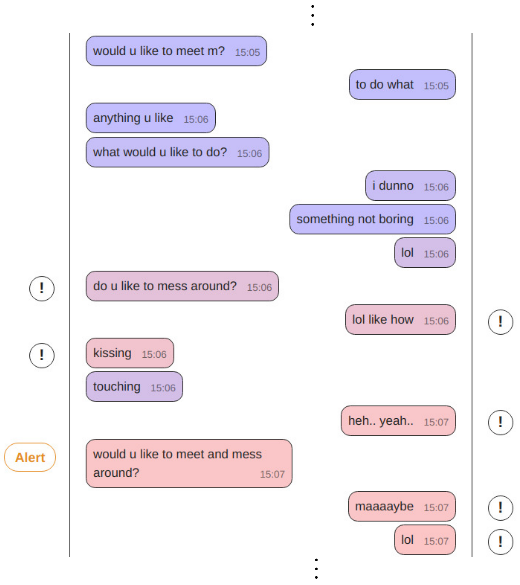
Early Detection of Sexual Predators in Chats

Matthias Vogt Ulf Leser Alan Akbik
Humboldt Universität zu Berlin
matthias.vogt@campus.tu-berlin.de
leser@informatik.hu-berlin.de
alan.akbik@hu-berlin.de

Abstract

An important risk that children face today is *online grooming*, where a so-called sexual predator establishes an emotional connection with a minor online with the objective of sexual abuse. Prior work has sought to automatically identify grooming chats, but only after an incidence has already happened in the context of legal prosecution. In this work, we instead investigate this problem from the point of view of prevention. We define and study the task of *early sexual predator detection* (eSPD) in chats, where the goal is to analyze a running chat from its beginning and predict grooming attempts as early and as accurately as possible. We survey existing datasets and their limitations regarding eSPD, and create a new dataset called PANC for more realistic evaluations. We present strong baselines built on BERT that also reach state-of-the-art results for conventional SPD. Finally, we consider coping with limited computational resources, as real-life applications require eSPD on mobile devices.



would u like to meet m? 15:05
to do what 15:05
anything u like 15:06
what would u like to do? 15:06
i dunno 15:06
something not boring 15:06
lol 15:06
do u like to mess around? 15:06
lol like how 15:06
kissing 15:06
touching 15:06
heh.. yeah.. 15:07
would u like to meet and mess around? 15:07
maaaaybe 15:07
lol 15:07

Figure 1: Visualization of chat messages and early sexual predator detection (eSPD). On each new message, the analysis is updated to reflect the level of risk. Finally, an alert is triggered as the risk-threshold is passed. Our goal is to detect such risk as early as possible. Note that real chats are much longer and can be non-contiguous conversations that span over weeks or months. Original source [3]

1 Introduction

Online grooming denotes the process where a so-called sexual predator establishes an emotional connection with a minor online to systematically solicit and exploit them for sexual purposes (Wachs et al., 2012). Online grooming is a major concern of public safety that, sadly, is rapidly growing. For instance, in England and Wales in the year to mid-2020, police recorded 5,083 offenses of *Sexual Communication with a Child* [1], an average of 14 offenses per day. In Germany, there were 2,632 recorded cases in 2020 where a child was sexually abused through internet communication technologies [2], an increase of 50 % to the previous year. As such crimes often go unreported or undetected, police-recorded incidents certainly do not fully reflect the real scale of the issue (Bowles and Keller, 2019; McGuire and Dowling, 2013).

The problem of detecting whether or not a child is being groomed by a predator is called *sexual predator detection* (SPD). Most previous approaches to SPD have cast this as the problem of identifying predatory authors in a corpus of segments of chats (Villatoro-Tello et al., 2012; Cardei and Rebedea, 2017). Other approaches interpreted it as a binary classification problem over segments of a chat (Ebrahimi et al., 2016), or the entire chat (Bours and Kulsrud, 2019). Approaches were evaluated mostly using data from the PAN shared task on sexual predator detection (Inches and Crestani, 2012). However, most prior work has

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing, pages 4985–4999
August 1–6, 2021. ©2021 Association for Computational Linguistics

viewed SPD from the point of view of forensics: they focused on identifying completed grooming chats in preparation for legal prosecution.

We believe that it is also important to study approaches that may *prevent* online grooming – as early as possible, i.e., during an ongoing chat. Ideally, the grooming process should be disrupted before it succeeds to protect children from harm. This task is non-trivial as the content of grooming chats changes over time: chats often start with the exchange of personal information and building of trust, a phase in which they are difficult to detect. In a second stage, predators further develop trust with their victims in a cycle of entrapment. They try to desensitize their victims to sexual topics, isolate them from others, and arrange meetings (Olson et al., 2007, p. 236). Even in this second stage, it is difficult to distinguish between grooming and consensual conversations between minors or adults. For this, a model needs to be able to detect discriminative features like a user talking about age difference, checking on the victim's relationship with their parents, isolating them from their support network, reframing sexual actions as appropriate and more (see Olson et al. (2007), pp. 234ff).

An example of arranging a meeting is shown in Figure 1. Here, an alert is triggered only late in the grooming process, when an in-person meeting is already explicitly being discussed. Ideally, such chats should be detected far sooner. However, the real-world consequences of a triggered eSPD alert can be considerable and may involve police actions. This means that false alerts should be avoided as much as possible. At the same time, false negatives must be avoided by all means as these could lead to a sexual assault. It is therefore as important as ethically difficult to find the best balance between the earliness of an alert and the certainty that an alert is justified.

1.1 Contributions

We introduce the task of *early sexual predator detection* (eSPD) in chats. We cast eSPD as an early risk detection problem in which chats are analyzed from the start and message by message, with the goal of raising warnings for chats early and accurately. Specifically, we make the following contributions:

- We introduce the problem of eSPD and formally define it.
- We survey available datasets, analyze their limitations, and build a new combined dataset called PANC as a best-effort for evaluating eSPD.
- We propose a task setup to evaluate eSPD, focusing on the trade-off between earliness and accuracy.
- We present strong baselines for eSPD using a two-tier approach. Our method (1) analyzes sliding windows of messages from an ongoing chat using BERT and (2) continuously classifies the sequence of the window classifications. We evaluate three different BERT language models, two of which work on mobile.
- We compare our models to previous research in conventional (i.e. "non-early") SPD settings and find that two of them outperform the current state of the art.
- We provide an extensive discussion of the limitations of our models and the available data.

We see our work as an important step to encourage more research into eSPD. To this end, we make our experimental setup, our baseline models, scripts for corpus processing, and the visualization tool for inspecting analyzed chats (used to generate Figure 1) publicly available[1]. We emphasize that we do not consider our models to be ready for use in real scenarios, which we discuss in depth in our Ethics Statement (see below).

2 Analysis of available datasets

Due to privacy and legal reasons, grooming chats are extremely difficult to obtain. We introduce the (few) known corpora of this kind and discuss their limitations, motivating the assembly of the PANC dataset we discuss in Section 3.

2.1 Original data sources

The main source of grooming chats used in SPD literature is the Perverted Justice Foundation (PJ) [10]. This organization used trained volunteers (decoys) posing as children in public chat rooms to help authorities convict sexual predators. They provide their chats with convicted predators for download but ceased their decoy operations in 2019. Nearly all prior work evaluates on datasets derived from PJ (McGhee et al., 2011; Gupta et al., 2012; Bogdanova et al., 2014; Meyer, 2015; Ebrahimi et al., 2016; Cardei and Rebedea, 2017; Pastor Ĺopez-Monroy et al., 2018).

[1] `early-sexual-predator-detection.gitlab.io`

To our knowledge, the only work using real grooming chats is Cheong et al. (2015) who used chats extracted from *MovieStarPlanet*, a massively multiplayer online game for children. Unfortunately, this corpus is not publicly available.

2.2 Corpora used for SPD

2.2.1 PAN12

The PAN Lab at the 2012 CLEF conference introduced a shared task on sexual predator identification [7]. The organizers created a large dataset which we call PAN12 using data from PJ. As non-grooming chats, they sampled from logs of IRC channels and of the chatting site Omegle [11]. These chats also include cybersex between consenting adults among non-predatory conversations, which makes distinguishing grooming chats especially difficult. They divided chats into *segments* whenever a conversation was interrupted for more than 25 minutes and filtered all segments with more than 150 messages. This results in a total of 222k segments, of which 2.58 % are grooming chats, through which the organizers try to mimic the distribution of grooming in actual online conversations. They are partitioned into train and test splits of a 30:70 ratio.

PAN12 has several limitations. All grooming chats stem from decoy operations and are not with actual victims, and the non-grooming chats are not with decoys. real. Most problematic for eSPD is the separation into relatively short, unordered segments, thus completely blurring the true timeline of a chat. This makes the data unsuitable for eSPD since we aim to detect predators as early as possible in potentially long-running chats.

2.2.2 VTPAN

Villatoro-Tello et al. (2012) found that filtering the PAN12 segments to only focus on the most important samples can lead to better model performance. They created a new dataset (VTPAN) by removing from PAN12 segments that have only one participant, less than 6 interactions per user, or long sequences of special characters (often depicting ASCII art). Many short segments which stem from predatory chats actually contain no predatory language, so a benefit of VTPAN is that many of these segments are filtered. The dataset is only 10% of the size of PAN12, and is also used in recent work on SPD (Escalante et al., 2016, 2017; Pastor Ĺopez-Monroy et al., 2018). Regarding eSPD, this dataset suffers from the same limitation as PAN12.

2.2.3 ChatCoder2

The *ChatCoder2* (CC2) corpus was created by McGhee et al. in 2011 and was later also used by other researchers (Basave et al., 2014). It contains 497 complete predator chats from PJ and was built mainly for studying the semantic segmentation of grooming chats. Accordingly, messages in 155 chats are also labeled as belonging to one of three phases: (1) exchange of *Personal Information*, (2) *Grooming*, and (3) *Approach* of the victim.

2.3 Limitations

In summary, we find that existing datasets suffer from limitations that make them difficult to use for training and evaluating eSPD. The commonly used datasets PAN12 and VTPAN only contain short, disjointed, and unordered chat segments. For eSPD, however, one needs to detect grooming in a continuous message stream, which is ordered and theoretically unbounded in length. Classifying segments only, we have no information about how early in the complete chat grooming is detected. Moreover, evaluating earliness within single segments would not be interesting as it is not interpretable and because they are so short. While CC2 does have full chat logs, it does not contain any negative samples. Our analysis thus motivates the assembly of the new PANC dataset as explained in the next section.

3 Early Sexual Predator Detection

In this section, we propose an evaluation setup for eSPD. We give a formal definition of the task followed by suitable evaluation metrics. Finally, we discuss how we use and combine existing SPD datasets to create PANC for the evaluation of eSPD.

3.1 Task definition

We interpret eSPD as an *early risk detection* problem (Losada et al., 2020). This means that we need to consider the earliness and the accuracy of warnings, continuously analyzing a chat after each new message. Formally:

Definition 1 (Message). A *message* is a string with a *time* and an *author*.

Definition 2 (Chat). A *chat* $C = (m_1, m_2, \dots)$ is a sequence of messages m_i where the time of messages is monotonically increasing. A *finite chat* is of the form $\widehat{C} = (m_1, \dots, m_n)$, where we say $\widehat{C}$ has a *length of* n. We call grooming chats *positive* and other chats *negative*. This is the *class* of a chat.

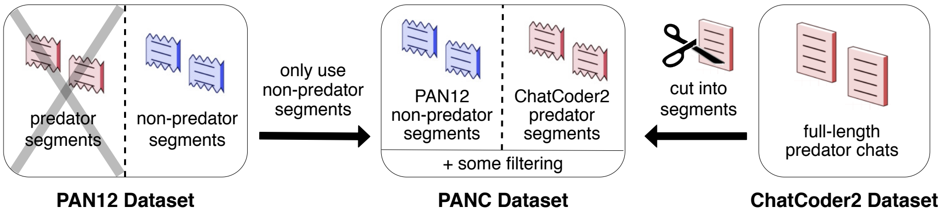


Figure 2: Creating PANC from PAN12 and CC2.

The length of real chats is potentially unbounded and keeps increasing, so regarding real chats as infinite is handy. We analyze chats after each new message, thus considering only finite prefixes for classification.

Definition 3 (Prefix). Let $C = (m_1, \dots, m_l, \dots)$ be a chat. We call $C(l) := (m_1, \dots, m_l)$ the prefix of C with length l.

Finally, we define eSPD as follows.

Definition 4 (eSPD). Let X_{Test} be a dataset of finite chats. For $C = (m_1, \dots, m_n) \in \mathrm{X}_{\text{Test}}$ and $l = 1, \dots, n$ increasing over time, an *eSPD system* decides for each l whether a warning for C should be raised or not by classifying $C(l)$. It stops as soon as a warning is raised, classifying C as grooming. If no warning is raised for all $l = 1, \dots, n$, it classifies C as non-grooming. Finally, *eSPD* is the problem of classifying all $C \in \mathrm{X}_{\text{Test}}$ as early and accurately as possible.

Note that this definition deliberately states that an eSPD system never classifies a chat as non-grooming as long as there are messages left (or the chat did not end, in a real-life setting), as it cannot know the future after the current prefix $C(l)$.

3.2 Evaluation metrics for eSPD

In eSPD, there are two desiderata between which a trade-off exists: (a) Raising alerts as early as possible, and (b) raising alerts as accurately as possible. Raising warnings early is good for (a), but hampers (b) as less data is available. Waiting longer with warning hurts (a), but most likely improves (b), as later decisions are based on more messages.

3.2.1 Accuracy of warnings

Accuracy metrics are most prominent in related work on detecting sexual predators (Pastor Ĺopez-Monroy et al., 2018; Escalante et al., 2017), i.e. "non-early" SPD. We report the established metrics of precision, recall, and F_1 for the grooming class.

3.2.2 Earliness of warnings

We call the number of messages that have been exchanged before a warning is raised the *warning latency*. We use latency-weighted F_1 (Sadeque et al., 2018) as a measure that accounts for both warning accuracy and warning latency. To calculate it, we first define a penalty for each warning latency $l \geq 1$ given by

$$\text{penalty}(l) := -1 + \frac{2}{1 + \exp(-p \cdot (l-1))}$$

where p determines how quickly the penalty should increase as latency increases. A warning after the first message receives 0 penalty and for increasing warning latency, the penalty approaches 1.

Now assume an eSPD system to produce a list latencies of warning latencies for all chats $C \in \mathrm{X}_{\text{Test}}$ where (1) C is positive, and (2) the system raises a warning for C. We define the overall speed of correct warnings as

$$\text{speed} := 1 - \text{median}\{\text{penalty}(l) \mid l \in \text{latencies}\}.$$

This metric is more interpretable than just using the mean or median warning latency, as it depends on the problem and the dataset at hand how good a median warning latency actually is. Finally, the latency-weighted F_1 is given by $F_{\text{latency}} := F_1 \cdot \text{speed}$. We generally consider an eSPD system A better than an eSPD system B when it reaches, for a given dataset, a higher F_{latency}; comparisons focusing more on speed or more on accuracy or searching for pareto-optimal solutions are also possible. Note that we, following Losada et al. (2019), compute the speed of warnings only for grooming chats classified as such. All other cases (false positives, false negatives, true negatives) are accounted for through the F_1 value.

3.3 The PANC dataset

Evaluating an eSPD system needs a corpus of chats, where each entire chat is annotated as grooming or not. Note that we do not require this annotation

	number of segments	positive segments	negative segments	% positive segments	Full-length pos. chats	pos. segment length *messages, words*	neg. segment length *messages, words*	length of full-length positive chats *messages, words*
PANC$_{\text{Train}}$	19,351	1,753	17,598	9.06 %	298	64 (±43), 289 (±218)	36 (±25), 173 (±1,385)	1,959 (±3,032), 8,730 (±12,223)
PANC$_{\text{Test}}$	13,159	1,426	11,733	10.84 %	199	65 (±43), 292 (±222)	36 (±26), 184 (±1,529)	2,248 (±3,141), 10,231 (±13,177)
PANC	32,510	3,179	29,331	9.78 %	497	64 (±43), 291 (±220)	36 (±25), 177 (±1,444)	2,075 (±3,079), 9,331 (±12,635)

Table 1: PANC overview. Segment/chat lengths are given through mean and standard deviation

on the message level, as what constitutes the first grooming message in a chat is highly subjective. Furthermore, eSPD based on supervised learning requires an annotated training corpus. Existing datasets cannot be directly used for this purpose, because they either consist only of unordered segments (VTPAN, PAN12), which hinders measuring speed, or only contain positive chats (CC2), which makes measuring F_1 impossible. Furthermore, the existing corpora all use PJ grooming chats and partly overlap.

To address these issues, we assembled PANC, an evaluation dataset for eSPD, by carefully combining selected parts from PAN12 and from CC2. The process is illustrated in Figure 2: The final corpus consists of (1) all positive full length chats from CC2 and (2) the negative segments of PAN12. We randomly split the corpus on this level at proportions 60:40 into train/test splits. Through (1), we can evaluate earliness. We cannot measure accuracy as defined above due to the lack of full-length negative chats. Instead, in the experiments, we will compute accuracy based on segments as an estimate of (2), for which we split the full-length grooming chats into segments. We filter all segments shorter than 6 messages, similar to VTPAN, and those longer than 150 messages (some of the latter were actually not filtered in PAN12, contrary to its original specification). Finally, we removed segments that are not between exactly two authors to make them comparable to CC2 chats. Statistics on the resulting corpus are given in Table 1.

Discussion. We consider PANC to be the first corpus suitable for realistic eSPD evaluations. Yet it still has limitations: First, the negative chats are not full-length chats but only segments. While this does not impact our earliness evaluation, it prevents the computation of true eSPD accuracy. Our proposed workaround is to replace chat accuracy with segment accuracy, although we do not know how well the latter approximates the former as we therein classify short segments which can stem from anywhere in a chat. An alternative would be to use a difference source for the negative chats; however, we decided on those from PAN12 as they also include "hard negative" cases (i.e. sexual conversations between consenting adults), which we believe gives more realism to our evaluation. Another limitation is that PANC only contains chats between exactly two authors, so our systems are not applicable in group chats. However, grooming is very rare in group chats as predators depend on their actions staying unnoticed.

4 Baseline approach: Two-Tier eSPD

We present a straightforward eSPD approach to demonstrate the validity of our task setup and to establish baselines for future works. It consists of two tiers of classification: (1) A local tier (Tier 1) that moves a sliding window over the messages of a chat and classifies them, and (2) a global tier (Tier 2) that decides after each window prediction whether to raise a warning or not based on the sequence of recent window predictions. The purpose of this architecture is to balance earliness and accuracy and especially to prevent single suspicious windows from triggering warnings.

4.1 Tier 1: Classifying sliding windows

For Tier 1, we use a standard approach in which we add a linear classifier to a pre-trained transformer model and fine-tune the entire architecture. It takes as input all messages in a given window and outputs a binary prediction. We evaluated different BERT models: BERT$_{\text{large}}$, BERT$_{\text{base}}$ (Devlin et al., 2018), and MobileBERT (Sun et al., 2020). Model parameters can be found in Appendix A. MobileBERT is a version of BERT$_{\text{large}}$ with smaller model size and faster inference, optimized for use on mobile devices.

Hyperparameters. Next to the choice of language model, the main hyperparameter of Tier 1 is the *window size*. It controls the number of messages that are input into the classifier.

4.2 Tier 2: Classifying chat prefixes

We use a simple approach for the problem of detecting a chat as grooming based on Tier-1 clas-

Approach	F_1	Precision	Recall	Speed	F_{latency}
$S_{\text{BERT-large}}$	0.88 ($\pm$0.05)	**0.88** ($\pm$0.03)	0.89 ($\pm$0.11)	0.75 ($\pm$0.17)	0.67 ($\pm$0.18)
$S_{\text{BERT-base}}$	**0.89** ($\pm$0.02)	0.82 ($\pm$0.04)	**0.96** ($\pm$0.01)	**0.91** ($\pm$0.02)	**0.81** ($\pm$0.03)
$S_{\text{MobileBERT}}$	0.80 ($\pm$0.04)	0.69 ($\pm$0.07)	0.95 ($\pm$0.01)	0.72 ($\pm$0.02)	0.58 ($\pm$0.02)

Table 2: Warning accuracy scores of our eSPD systems on PANC (as mean and standard deviation)

sification results over a series of windows. After every window classification, we consider the count of positively classified windows within the last 10 windows. If this value exceeds a pre-defined threshold called *skepticism* $s \in \{1, \ldots, 10\}$, the chat is classified as grooming.

Hyperparameters. The only hyperparameter of Tier-2 is thus *skepticism* which controls the earliness/accuracy tradeoff.

5 Evaluation

We evaluate our baseline approach in our eSPD task setup using the proposed metrics for warning earliness, accuracy, and F_{latency}. We compare three different eSPD systems: $S_{\text{BERT-large}}$, $S_{\text{BERT-base}}$, and $S_{\text{MobileBERT}}$, which use the respective transformer models as described above as the Tier-1 classifier. We use a window size of 50 and a skepticism of 5; an evaluation of the impact of the skepticism parameter can also be found below. We fine-tune each of our BERT models on PANC and VTPAN. As the results of fine-tuning BERT models often vary heavily based on the random seed used (Dodge et al., 2020), we repeat this process three times. In the evaluation, we always report the mean of the resulting measures together with standard deviation. We fine-tune BERT$_{\text{base}}$ and MobileBERT using the TensorFlow Lite Model Maker [8] Library and BERT$_{\text{large}}$ using Flair [9] (Akbik et al., 2019).

5.1 Experimental results

An overview of evaluation results for our three model variants is given in Table 2. To compute the F_{latency} of warnings, we measured their F_1 score for segments, while speed is based on full positive chats (see Section 3).

Evaluating earliness in isolation. Figure 3 shows violin plots of the distribution of warning latencies for the three systems for all predator chats from PANC$_{\text{Test}}$, based on the means over three runs. The systems $S_{\text{BERT-large}}$ and $S_{\text{MobileBERT}}$ have similar performance while $S_{\text{BERT-base}}$ outperforms both. Its median warning latency is roughly 30 messages lower compared to the other systems. Moreover, $S_{\text{BERT-base}}$ exhibits much less variance in warning latency than the other two models. An explanation of the somewhat surprising scores of $S_{\text{BERT-large}}$ is that one of the three runs of this model led to significantly worse results than the other runs. As a consequence, the standard deviation of this model is also much higher than for the other two models.

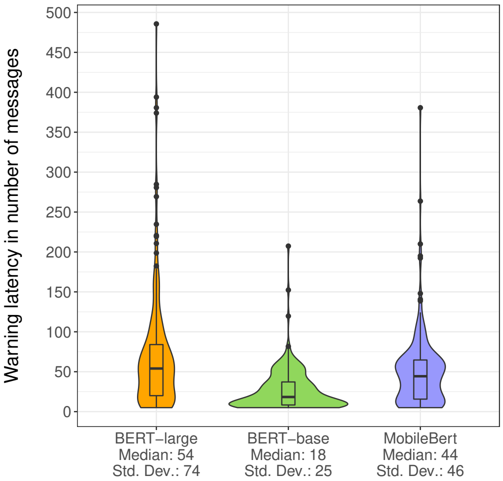


Figure 3: Warning latency distributions of our systems for the full-length predator chats in PANC$_{\text{Test}}$.

Interpreting and penalizing warning latency. To calculate F_{latency}, we need to set the parameter p which controls the penalty that is assigned to a given warning latency. However, when inspecting the full-length predator chats, we noticed that the number of messages before a chat gets suspicious varies heavily, and there is no "typical" value for this, which makes setting p difficult. We believe that it would be better to not set p globally but on a chat by chat basis, which could be done in future work. Conventionally (Sadeque et al., 2018; Losada et al., 2019), p is set such that the penalty is 0.5 at the median length of chats. But for our full-length predator chats, this would be 1,055 messages which we think is way too late to raise a warning. Ultimately, we decided to set p with help from the message labels from CC2. We set p such that the penalty is 0.5 when about 20 grooming messages are exchanged. In median for the labeled CC2 chats, this is 90 messages, so we set $p = \ln(3)/(90 - 1) \approx 0.0123$. However, the standard deviation for this is about 200 messages.

Best baseline approach. As Table 2 shows, overall results differ whether one considers only F_1 or F_{latency}: Considering F_1, $S_{\text{BERT-large}}$ and $S_{\text{BERT-base}}$ have similar performance and both outperform $S_{\text{MobileBERT}}$. However, when considering speed, BERT$_{\text{base}}$ significantly outperforms the other models. One of the BERT$_{\text{large}}$ runs only scored a speed of 0.55 , which is why the mean speed is unexpectedly low and the standard deviation is high. In F_{latency}, $S_{\text{BERT-base}}$ outperforms $S_{\text{BERT-large}}$ by 0.14 which again outperforms $S_{\text{MobileBERT}}$ by 0.09 .

Impact of skepticism. The skepticism hyperparameter s controls the propensity of the Tier-2 classifier to raise warnings and can thus be seen as the central knob to tune the earliness/accuracy trade-off for our approach. We would expect that being more skeptical leads to a lower recall, higher precision, and higher latency of warnings. To confirm this, we evaluate each of our eSPD systems on PANC for each skepticism $s = 1, \ \ldots \ , 10$ and note precision, recall, and speed of warnings depending on skepticism. Here, the speed of warnings is calculated as explained in Section 3.2.2.

In Figure 4, we plot the concrete accuracy and speed metrics of our eSPD systems, depending on the skepticism of the Tier-2 classifiers. For all of our systems, we indeed find that as skepticism increases, precision increases as well, while recall and speed are decreasing. Moreover, the F_{latency} of our detectors does not significantly change as long as s is in a medium range of $\{3, 4, 5, 6, 7\}$, except for $S_{\text{BERT-large}}$, but here the standard deviation of F_{latency} is so high that no clear correlation exists.

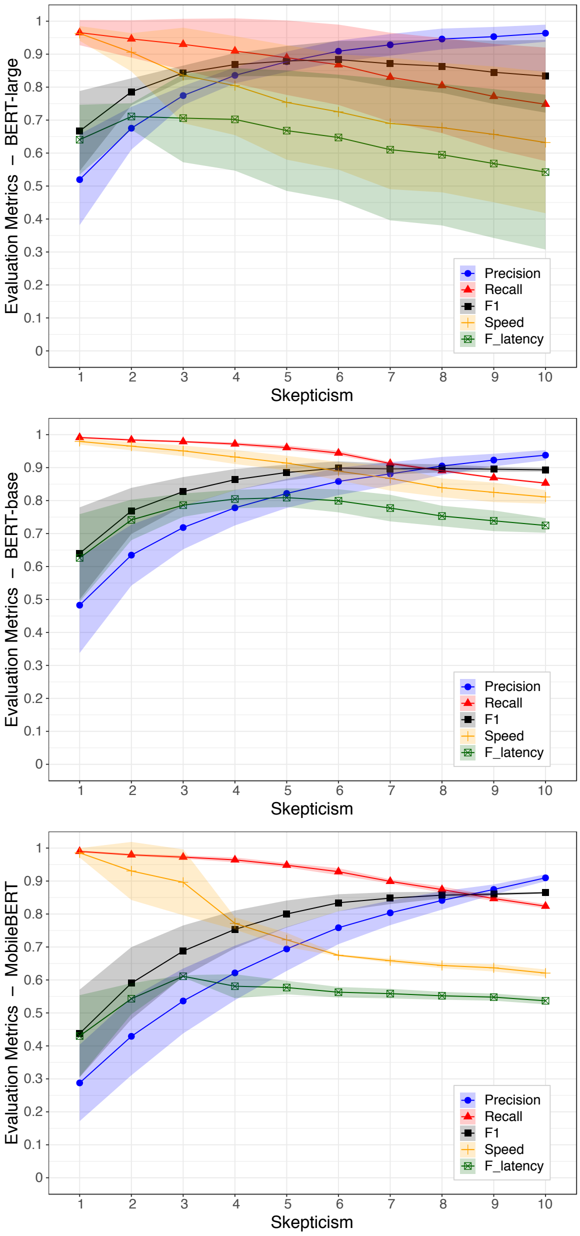


Figure 4: Impact of master classifiers skepticism s for our eSPD systems $S_{\text{BERT-large}}$, $S_{\text{BERT-base}}$, and $S_{\text{MobileBERT}}$. Dots and lines are the mean across different runs and the shaded area is the standard deviation

5.2 Comparison to conventional SPD

To get a better understanding of the accuracy of our proposed baseline approach, we also employ it in a conventional SPD setting. This allows us to compare against the state-of-the-art approaches by Escalante et al. (2017) and Pastor Ĺopez-Monroy et al. (2018).

Evaluation setup. For this comparison, we replicate their evaluation setting in which they classify segments on VTPAN by considering increasing fractions of each segment as measured by the number of characters. They evaluate their SPD accuracy after 10%, 20%, . . . , 100% of all characters of a segment where only whole words are included. As classification is not message-by-message, we only use our Tier-1 classifiers in this setting. Note that evaluating accuracy as a function of fraction of a segment also may be interpreted as earliness, though in a very different sense than proposed for eSPD in this paper, because segments are much shorter than chats and may be from anywhere within a chat.

New state of the art on SPD. Figure 5 summarizes the results of this comparison. Notably, even the MobileBERT model is competitive with previous works in spite of being much less resource hungry. Both other models outperform previous works for all settings. The difference in performance is especially large for small segment prefixes and decreases with increasing availability of information. For 10 % of information, BERT$_{\text{large}}$ outperforms the SOTA by as much as 8 % in F_1. A complete list of the F_1 values is given in Appendix B.

Discussion. We believe that improvements primarily stem from our usage of BERT, which previously had not been applied to SPD. The implementations of previous approaches are not openly available, so we cannot directly compare example inputs. But prior work uses document representations where words are considered irrespective of their context. Thus, we believe that these approaches are mostly able to detect grooming attempts that use specific words, for instance those with a sexual connotation. A BERT-style transformer model on the other hand may be able to better distinguish whether the overall context in which words are used is a grooming context and identify attempts that use more indirect language such as innuendo.

6 Discussion

We discuss several issues that must be considered before planning to apply an algorithm like the ones presented in this work in practice.

6.1 Language in (non-)grooming chats

A critical question is how representative PANC is of real grooming chats. Chiang and Grant (2019, p. 693) and Schneevogt et al. (2018), suggest that the PJ chats created by adult decoy volunteers instead of actual child victims (see Section 2.1) may not truly represent real grooming chats. Specifically, they found that they are missing themes of forceful persuasion or extortion of victims, which is present in real grooming chats. Furthermore, youth language changes very fast over the years; as our corpus is from 2012, it is questionable how well it would represent current chats. For instance, it does not contain any emojis. Another issue is the lack of deep relationships in our non-grooming chats. Among those, the only chats with personal or intimate conversations are from Omegle. This is a platform that invites cybersex, for example, but users do not have a strong personal relationship as they randomly meet (only) online. An example of how the lack of such chats might lead to false positives is shown in Appendix C.

6.2 Lack of complete negative chats

Due to the lack of publicly available datasets, we could not test our models on complete negative chats. This has implications: We had to resort to measuring accuracy at the segment level, and we cannot provide concrete estimates on warning accuracy for such chats. However, we consider our results on negative segments to be promising.

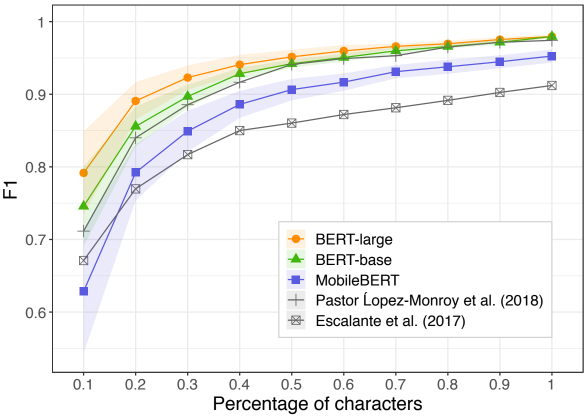


Figure 5: Our BERT models vs. SOTA on VTPAN classifying 10%, 20%, ... of characters of segments. Plotted lines represent the mean results of three runs and the shaded regions represent standard deviation.

6.3 Segment versus window classification

Our Tier-1 classifiers are trained on segments of a chat, created by a specific partitioning of the sequence of messages. However, during eSPD we apply them to windows of the last 50 messages, which may exhibit different properties than the predefined segments. For instance, as segments are separated by lengthy breaks in the conversation, they often begin with greetings – which is not the case for our windows. Such differences may confuse our models and lead to sequences of wrong window classifications, an effect we counteract through the Tier-2 classifier.

6.4 Use of additional information

While we consider only chat messages as information to detect grooming attempts, real-world applications might also have additional data available. For instance, in social media, users are often required to state their age when they create their profile. Such data could be very helpful for eSPD. However, we caution that profile information may not be reliable as it is typically not verified and therefore easy to fake – and it is common for predators to use fake information.

7 Related work

Online grooming is a real and pressing problem faced by any chat system open to children. Accordingly, social media sites and games often use automated grooming detection systems (Bowles and

Keller, 2019). For example, YouTube applies NLP to detect predatory messages in video comments and livestream chats followed by human verification (IICSA and Canegallo, 2019, p. 63, ll. 10–25). Microsoft uses a similar approach for XBOX Live and Skype chat [6] and also licenses their software to other service providers free of charge (Patel, 2020). Their obvious advantage over academic research is the access to much larger datasets. However, these solutions are server-based and cannot be applied for end-to-end encrypted chats. Many parents also resort to using parental control apps, some of which send children's chats to external servers for analysis, which is a privacy concern. Because of these reasons, there is a need for eSPD systems even on mobile devices.

In academia, eSPD so far has seen comparably little research despite its high societal importance, probably due to the difficulties of obtaining appropriate datasets. Villatoro-Tello et al. (2012) was the winning team of the first problem of the PAN12 competition, which was the identification of the predatory authors of the PAN12 segments. They approached the problem by first predicting segments as grooming or not and then distinguishing victim from predator. This two-step method was refined by Cardei and Rebedea (2017) who additionally used behavioral features, such as the number of questions asked, achieving an $F_{0.5}$ of 0.934 for segment classification on a subset of PAN12$_{\text{Test}}$. Bours and Kulsrud (2019) studied the same problem and included an analysis of early segment classification, i.e., an attempt to find predators early within a segment. They explored their method also by applying it to 10 full-length PJ chats, which could be seen as the first instance of eSPD we are aware of.

Early text classification. To our knowledge, Escalante et al. (2016) was the first work to approach SPD from an early text classification perspective, but restricted their analysis to the segment level. Their results were improved in Escalante et al. (2017) using profile-based representations, where documents are represented as normalized sums of vector representations of words. The best results so far for early segment classification were achieved by Pastor Ĺopez-Monroy et al. (2018) using a Multi-Resolution Representation (MulR) for documents to cope equally well with longer and shorter segments. We compared to the results of the latter two works in Section 5.2 and found that our approach outperforms both. Note that we are not aware of any previous work employing transformers for SPD.

Early time series classification. An interesting perspective on our Tier 2 is that it actually solves an early time series classification (eTSC) problem, for which there exist several mature approaches, e.g. TEASER (Schäfer and Leser, 2020) or ECTS (Xing et al., 2012). However, there exists a key difference that prevents us from using such methods directly: An eSPD System never classifies a chat as non-grooming as long as there are still messages left (or expected), while an eTSC system at some stage might decide that it is safe to stop controlling the chat (Loyola et al., 2018). This opens the door to malicious attacks by using long and harmless openings in grooming attempts. We nevertheless believe exploring ways to adapt eTSC to eSPD to be an interesting avenue for future research.

8 Conclusion

We defined the problem of early sexual predator detection (eSPD) in online chats and proposed an evaluation setup for this task. To this end, we assembled the PANC dataset, which, albeit having clear limitations, in our mind is the currently best effort possible with the data available. We also showed that a baseline built on current BERT-based language models achieves strong results on this dataset, and beats previous methods in related settings. Notably, results are only modestly impacted for models that can run on mobile devices. We discussed open issues in our data and evaluation setup that must be studied carefully in future work before eSPD systems could go live (and expand on this discussion in Appendix D). We hope that making our task setup accessible to the research community will encourage more research into the highly important topic of early sexual predator detection.

Acknowledgments

We would like to thank Dr. Hugo Jair Escalante and Dr. Esaú Villatoro-Tello for providing us with VTPAN and allowing us to publish the means to recreate it, as well as Professor April Edwards for providing us with CC2. We thank the Institute of Sexology and Sexual Medicine at Charité – Universitätsmedizin Berlin for introducing us to the problem of online grooming and the need for automatic solutions, and fruitful discussions. Finally, we thank the TensorFlow Lite Support Team and specifically Chen Cen for creating a workaround that enables our BERT models to work on mobile.

Ethics Statement

Early sexual predator detection is a highly sensitive topic which calls for a proper discussion of potential implications of such research, the datasets being used, and the readiness of eSPD models. There are potentially high stakes for any subject whose chats are analyzed by eSPD systems. Any application of eSPD in running chat systems would incur interaction with vulnerable populations (minors) which must be firmly protected. False-negative, as well as false-positive predictions, may have severe implications for the falsely alleged chat partner or the erroneously unprotected child, respectively. Online grooming is forbidden by law in many countries, as are the establishment of sexual relationships of any kind to children. In many countries, including Germany, already obtaining logs of chat content with sexual content involving children is forbidden, which makes acquisition or usage of real data impossible outside criminal investigations. At the same time, online grooming does happen now, and in many instances, making research into ways to prevent or at least diminish it important.

Datasets. For this study, we did not create any new data or perform any experiments with human beings. According to European regulations, such research does not require an ethics vote from an institutional review board. Instead, we performed specific filtering and combination of data from the two datasets PAN12 and ChatCoder2 (CC2), which are available on request to their authors, and have been extensively used in the literature.

The creators of PAN12 anonymized the data by removing usernames and email addresses to avoid the identification of users. This makes PAN12 compatible with European regulations that permit the exchange of carefully anonymized data. The CC2 chats stem from PJ and are with offenders who were prosecuted in court and adult decoys posing as children. Thus, they contain no conversations with minors or victims, which makes CC2 compatible with the above-mentioned regulations against possession and usage of any real chat logs involving sexual content with children.

Readiness of eSPD models. Real-world applications already use automatic systems to support detection of grooming in chats (Patel, 2020; Bowles and Keller, 2019), yet no details about their measured performance and internal functioning are known to us. However, we do not consider the models and methods presented in this paper as ready for production systems. We already discussed some of their technical limitations in Section 6. On top of these, we believe that any eSPD system must be carefully adapted to any concrete chat system and continuously retrained and monitored to be able to pick up specific styles of communication and how they change over time. Additionally, any system applying eSPD must take an ethically highly difficult decision regarding the trade-off between the two immanent desiderata for eSPD systems: the earliness of warnings and their accuracy. Perfectly achieving both, i.e., performing only correct classifications after the very first message, is impossible. In this research paper, we studied the impact of our *skepticism* factor which controls this trade-off. The concrete setting of this (or a similar) parameter in a real application must depend on an independent and careful assessment of consequences of false positive and false negative alarms. This decision must take the respective circumstances into account and requires an application-specific ethical assessment of its own, including options of monitoring by human professionals as discussed in Appendix D.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Amparo Cano Basave, Miriam Fernández, and Harith Alani. 2014. Detecting child grooming behaviour patterns on social media. In *SociInfo 2014: The 6th International Conference on Social Informatics*.

Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. 2014. Exploring high-level features for detecting cyberpedophilia. *Computer Speech and Language*.

Patrick Bours and Halvor Kulsrud. 2019. Detection of Cyber Grooming in Online Conversation. *2019 IEEE International Workshop on Information Forensics and Security, WIFS 2019*, pages 9–12.

Nellie Bowles and Michael H. Keller. 2019. Video games and online chats are 'hunting grounds‘ for sexual predators. The New York Times, `nytimes.com/interactive/2019/12/07/us/video-games-child-sex-abuse.html`. (Accessed on 2020/07/15).

Claudia Cardei and Traian Rebedea. 2017. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Natural Language Engineering*, 23(4):589–616.

Yun Gyung Cheong, Alaina K. Jensen, Elin Rut Gudnadottir, Byung Chull Bae, and Julian Togelius. 2015. Detecting Predatory Behavior in Game Chats. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):220–232.

Emily Chiang and Tim Grant. 2019. Deceptive identity performance: Offender moves and multiple identities in online child abuse conversations. *Applied Linguistics*, 40(4):675–698.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv preprint arXiv:2002.06305*.

Mohammadreza Ebrahimi, Ching Y. Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep Convolutional Neural Networks. *Digital Investigation*.

Hugo Jair Escalante, Manuel Montes y Gomez, Luis Villasenor, and Marcelo Luis Errecalde. 2016. Early text classification: a Naïve solution. *Proceedings of NAACL-HLT 2016, 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 91–99.

Hugo Jair Escalante, Esaú Villatoro-Tello, Sara E. Garza, A. Pastor López-Monroy, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. 2017. Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111.

Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. 2012. Characterizing Pedophile Conversations on the Internet using Online Grooming. *arXiv preprint arXiv:1208.4324*.

IICSA and Kristie Canegallo. 2019. IICSA Inquiry - Internet Hearing. `iicsa.org.uk/key-documents/11479/view/open-session-transcript-16-may-2019.pdf`. Video transcripts are also available at `youtube.com/playlist?list=PLQrDHIqFcNWFQRXe4pIAK4pnTEWn16IMR`.

Giacomo Inches and Fabio Crestani. 2012. Overview of the International Sexual Predator Identification Competition at PAN-2012. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 30.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15.

David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019 Early Risk Prediction on the Internet. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of eRisk 2020: Early Risk Prediction on the Internet (Extended Overview). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12260 LNCS:272–287.

Juan Martín Loyola, Marcelo Luis Errecalde, Hugo Jair Escalante, and Manuel Montes y Gomez. 2018. Learning when to classify for early text classification. *Communications in Computer and Information Science*, 790:24–34.

India McGhee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra McBride, and Emma Jakubowski. 2011. Learning to identify Internet sexual predation. *International Journal of Electronic Commerce*.

Mike McGuire and Samantha Dowling. 2013. Cyber crime: A review of the evidence. UK Home Office, Home Office Research Report 75. `gov.uk/government/publications/cyber-crime-a-review-of-the-evidence`.

Maxime Meyer. 2015. Machine learning to detect online grooming. Master's thesis, Uppsala Universitet.

Loreen N. Olson, Joy L. Daggs, Barbara L. Ellevold, and Teddy K.K. Rogers. 2007. Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251.

A. Pastor Ĺopez-Monroy, Fabio A. Gonźalez, Manuel Montes-Y-Ǵomez, Hugo Jair Escalante, and Thamar Solorio. 2018. Early text classification using multi-resolution concept representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

Priti Patel. 2020. New ai technique to block online child grooming launched. `gov.uk/government/news/new-ai-technique-to-block-online-child-grooming-launched`. (Accessed on 2020/07/19).

Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the latency of depression detection in social media. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, 2018-Febua:495–503.

Patrick Schäfer and Ulf Leser. 2020. TEASER: early and accurate time series classification. *Data Mining and Knowledge Discovery*.

Daniela Schneevogt, Emily Chiang, and Timothy Grant. 2018. Do Perverted Justice chat logs contain examples of Overt Persuasion and Sexual Extortion? A Research Note responding to Chiang and Grant (2017, 2018). *Language and Law = Linguagem e Direito*, 5(1):97–102.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: Task-Agnostic Compression of BERT by Progressive Knowledge Transfer. *arXiv preprint arXiv:2004.02984*.

Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. 2012. A Two-step Approach for Effective Detection of Misbehaving Users in Chats - Notebook for PAN at CLEF 2012. *CLEF (Notebook Papers/Labs/Workshop)*.

Sebastian Wachs, Karsten D Wolf, and Ching-Ching Pan. 2012. Cybergrooming: risk factors, coping strategies and associations with cyberbullying. *Psicothema*, 24(4):628–33.

Zhengzheng Xing, Jian Pei, and Philip S. Yu. 2012. Early classification on time series. *Knowledge and Information Systems*.

Links

[1] Instagram most recorded platform used in child grooming crimes during lockdown — NSPCC
nspcc.org.uk/about-us/news-opinion/2020/instagram-grooming-crimes-children-lockdown/

[2] BKA – PKS Tabellen – Thematische Gliederung – PKS 2020 Bund – Falltabellen; T05 Grundtabelle - Straftaten mit Tatmittel „Internet" – Fallentwicklung (V1.0); See "Einwirken auf Kinder § 176 Abs. 4 Nr. 3 und 4 StGB"
bka.de/DE/AktuelleInformationen/StatistikenLagebilder/PolizeilicheKriminalstatistik/PKS2020/PKSTabellen/BundFalltabellen/bundfalltabellen.html

[3] Full log of the chat gjk1352 from PJ.
perverted-justice.com/?archive=gjk1352

[4] Full log of the chat bloodlineofhate from PJ.
perverted-justice.com/?archive=bloodlineofhate

[5] Full log of the chat ich_bin_der_eggman_67 from PJ.
perverted-justice.com/?archive=ich_bin_der_eggman_67

[6] Microsoft shares new technique to address online grooming of children for sexual purposes - Microsoft On the Issues
blogs.microsoft.com/on-the-issues/2020/01/09/artemis-online-grooming-detection/

[7] PAN is a series of scientific events and shared tasks on digital text forensics and stylometry.
pan.webis.de

[8] Text classification with TensorFlow Lite Model Maker
tensorflow.org/lite/tutorials/model_maker_text_classification

[9] flairNLP/flair: A very simple framework for state-of-the-art Natural Language Processing (NLP)
github.com/flairNLP/flair

[10] Perverted Justice Foundation Homepage
perverted-justice.com

[11] Omegle: Talk to strangers!
omegle.com

[12] PAN12 Deception Detection: Sexual Predator Identification — Zenodo
zenodo.org/record/3713280

[13] ChatCoder Data page
chatcoder.com/data.html

[14] CyberTipline
missingkids.org/gethelpnow/cybertipline

[15] Facebook encryption plans will hit fight against child abuse, warns Patel
theguardian.com/society/2021/apr/19/priti-patel-says-tech-companies-have-moral-duty-to-safeguard-children

All accessed June 1, 2021.

Appendix

A Models and training

The parameters for our BERT models can be found in Table 3. The hyperparameters we used for fine-tuning our BERT models are listed in Table 4. We fine-tuned the models on a high-end compute server which has an NVIDIA Tesla V100 GPU with 32GB of RAM, an Intel Xeon 6254 Processor, and 756GB of RAM.

B Specific Evaluation results for comparison with SOTA

Table 5 gives the specific F_1 scores for the evaluation in Section 5.2.

C Examples of grooming chats without predatory language that are classified as predatory

Next to cybersex chats, an important possibility for false-positive warnings in practice are chats between lovers. Such chats are most likely very rare among the negative PAN12 segments, which we use for PANC. In Figure 6, we can see two excerpts from positive chats for which $S_{\text{BERT-large}}$ raises warnings. In our opinion, out of context, the excerpts could just as well occur in a regular chat between lovers, so they should not be classified as grooming by themselves. This is an example of a feature that is discriminative in our datasets but not in reality.

D Further discussion: Scenarios for applications of eSPD systems

We see two main operational modes in which eSPD systems as presented in this paper could be deployed. Chats may either be analyzed centrally, i.e. at the messenger's server, or decentrally, i.e. at the chat clients. These modes lead to very different situations regarding the earliness/accuracy trade-off.

Server-side systems. In most systems, chats are stored on a server of the chat provider. This enables a *hybrid* setup that combines automatic predictions with manual verification by experts: eSPD systems would be used to flag suspicious chats at scale which are then referred to trained professionals. Only if professionals agree, proper actions would be taken, like stopping the chat, notifying certain persons, referring to appropriate institutions like the `CyberTipline` [14], or informing the police. Such a hybrid approach reduces the danger of false alarms. For example, YouTube handles live-stream chats with such an approach, as stated in IICSA and Canegallo (2019, p. 63, ll. 10–25).

However, even in a hybrid setup, we encounter ethical questions regarding the trade-off between earliness and accuracy of warnings. If the eSPD system prioritizes earliness and thus raises many warnings, it might result in a flood of warnings that can quickly overwhelm moderators. Moreover, mass moderation of user chats could raise privacy concerns. On the other hand, if the system prioritizes accuracy, this may lead to a failure to prevent sexual assaults of minors. Finding the specific balance for a given application requires careful ethical considerations whose reasoning should be made transparent to the users, and in case of minors especially to the parents. To use a messaging application, users should have to give informed consent to the system-specific regulations, the modes of control and moderation, and the potential risks of the implemented strategies.

Client-side systems. As many messaging (and chatting) systems are moving toward end-to-end encryption [15], the previously described mode of centralized application of eSPD becomes increasingly infeasible, as neither moderators nor software are able to decrypt the chats once they left the device of the chatting persons. In this case, eSPD systems can only be deployed on the chat client, which is in most cases a smartphone. They could be installed separately from the client, or be already integrated into the client. The latter could result in warnings being created both at the side of the child and at the side of the potential predator; both cases must be analyzed carefully. Note that during installation, the software is not able to control whether it is being installed on the smartphone of an adult or of a minor.

On the child side, systems could be configured to (1) send alerts to the parents of the minor, (2) directly alerting the minor, or (3) both. Option (2) is beneficial for the privacy of the minor, but places a higher responsibility on them to adequately deal with warnings. A grooming alert would have to be communicated very carefully to not be traumatizing. In all of the above cases, children could also mistakenly assume that the eSPD system is a bullet-proof "safety net" that allows them to be less careful when chatting online. A missing alert

Model	Version	Max-Seq.	L	H	A	Params.	Mobile	Model Size (quantized)	Inference Latency (mobile)
BERT$_{\text{large}}$	uncased	512	24	1024	16	336 M	✗	1,300 MB	20 ms
BERT$_{\text{base}}$	uncased	512	12	768	12	110 M	✓	419 MB (106 MB)	2,700 ms (5,410 ms)
MobileBERT	uncased	512	24	128	4	25 M	✓	95 MB (25 MB)	800 ms (1,907 ms)

Table 3: Overview of the BERT models we used for our tier-1 classifiers. The models have L Layers, Hidden size H and A Attention Heads. Inference latency shows the average desktop/mobile inference latency. For BERT$_{\text{base}}$ and MobileBERT, we ran the converted TensorFlow Lite models on desktop/mobile, which is still experimental and not well optimized yet. The mobile inference latencies are for the quantized versions of the models which we ran on a Sony Xperia XZ1 compact which has an Octa-core CPU (4x2.45 GHz Kryo & 4x1.9 GHz Kryo).

Hyperparameter	Value
Optimizer	Adam (Kingma and Ba, 2015)
Loss function	Crossentropy
Epochs	3
Mini batch size	16
Initial learning rate	$3 \cdot 10^{-5}$ for BERT$_{\text{large}}$, else $5 \cdot 10^{-3}$

Table 4: Overview of training hyperparameters

could be interpreted as that a chat is safe, no matter what is being communicated, which would actually reduce the safety of the child. Options (1) and (3) create greater safety for the minor, but at the risk that parents are sometimes falsely warned in situations where no online grooming takes place, which could quickly result in psychologically delicate situations. Parents are not trained professionals, as moderators are, which increases the chances of misunderstanding warnings. One can imagine that uninformed and very cautious parents call the police in any case of a warning without any further checks, which in case of false alarms would lead to wrong allegations, psychological stress, and societal stigma on part of the accused.

We should further consider that a predator could use an eSPD system and monitor its assessment of the ongoing chat to anticipate warnings. This could signal the predator to change wording and language to circumvent detection. One could even imagine systems where the predator can check if sending a specific message would trigger a warning (by running a second, parallel yet faked chat). While messages by the victim could also trigger warnings, the predator could still use the method to make detection much less likely, and possibly to learn how to use language to elude the system. In any case, to avoid predators finding ways to circumvent detection, the specific eSPD system used by the application should not be made available separately. Users should also not be able to see the current "risk level" of a chat or to control the sensitivity of warnings.

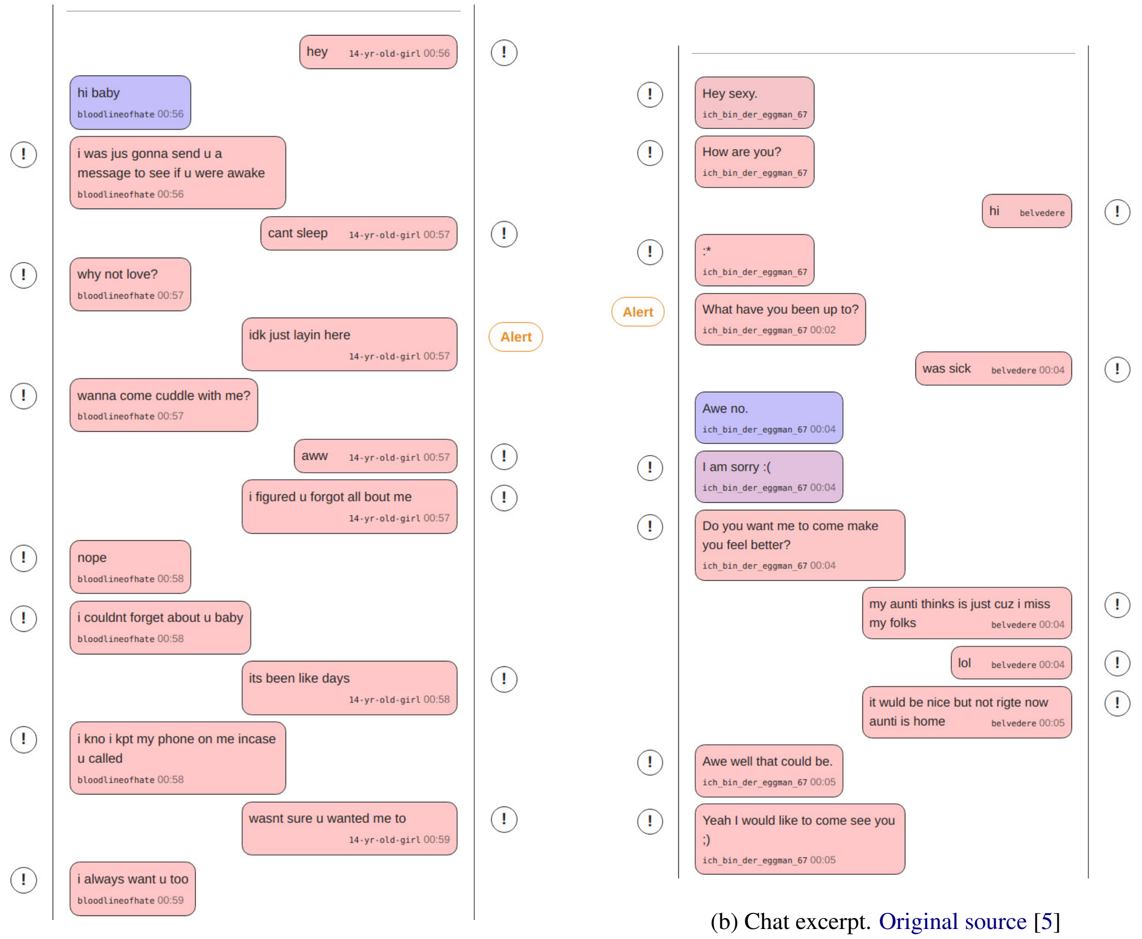
Overall, client-side systems thus face challenges in how and to whom to raise warnings. Warnings should on the one hand be disruptive enough to be taken seriously by the user while at the other hand clearly communicating that a warning is only an estimation and therefore does not establish guilt. Users might also be given the option to disable alerts for certain contacts whom they trust to reduce the number of false alerts. For systems that raise warnings to a minor's parents, it would be important to include clear messaging to the parent that eSPD systems are not perfect and may both raise false warnings as well as miss actual grooming attempts. The parent should be clearly advised that an eSPD system offers only partial protection and that it is still important to teach their children how to identify a dangerous chat themselves.

E Supplementary Material

Our evaluation setup, dataset preprocessing code, trained models, and chat visualization software can be found at `early-sexual-predator-detection.gitlab.io`. We are not allowed to distribute the PAN12 and CC2 datasets which are available on request to the respective dataset's original authors (see [12, 13]).

Approach	10% of characters	20% of characters	30% of characters	40% of characters	50% of characters
$\text{BERT}_{\text{large}}$	**0.7916** (± 0.0574)	**0.8908** (± 0.0261)	**0.9230** (± 0.0168)	**0.9408** (± 0.0135)	**0.9515** (± 0.0098)
$\text{BERT}_{\text{base}}$	0.7457 (± 0.0551)	0.8558 (± 0.0275)	0.8969 (± 0.0162)	0.9284 (± 0.0082)	0.9421 (± 0.0056)
MobileBERT	0.6285 (± 0.0854)	0.7923 (± 0.0389)	0.8492 (± 0.0283)	0.8860 (± 0.0187)	0.9064 (± 0.0148)
Pastor Ĺopez-Monroy et al. (2018)	0.7115	0.8400	0.8856	0.9166	0.9411
Escalante et al. (2017)	0.6710	0.7697	0.8169	0.8500	0.8603
Approach	60% of characters	70% of characters	80% of characters	90% of characters	100% of characters
$\text{BERT}_{\text{large}}$	**0.9596** (± 0.0085)	**0.9660** (± 0.0035)	**0.9696** (± 0.0044)	**0.9754** (± 0.0034)	**0.9796** (± 0.0027)
$\text{BERT}_{\text{base}}$	0.9507 (± 0.0057)	0.9598 (± 0.0049)	0.9657 (± 0.0026)	0.9716 (± 0.0033)	0.9794 (± 0.0014)
MobileBERT	0.9167 (± 0.0120)	0.9311 (± 0.0092)	0.9379 (± 0.0097)	0.9448 (± 0.0091)	0.9527 (± 0.0091)
Pastor Ĺopez-Monroy et al. (2018)	0.9492	0.9531	0.9650	0.9716	0.9743
Escalante et al. (2017)	0.8721	0.8814	0.8916	0.9025	0.9121

Table 5: Specific F_1 scores as mean and standard deviation for the evaluation in Section 5.2



hey 14-yr-old-girl 00:56

hi baby
bloodlineofhate 00:56

i was jus gonna send u a message to see if u were awake
bloodlineofhate 00:56

cant sleep 14-yr-old-girl 00:57

why not love?
bloodlineofhate 00:57

idk just layin here
14-yr-old-girl 00:57

wanna come cuddle with me?
bloodlineofhate 00:57

aww 14-yr-old-girl 00:57

i figured u forgot all bout me
14-yr-old-girl 00:57

nope
bloodlineofhate 00:58

i couldnt forget about u baby
bloodlineofhate 00:58

its been like days
14-yr-old-girl 00:58

i kno i kpt my phone on me incase u called
bloodlineofhate 00:58

wasnt sure u wanted me to
14-yr-old-girl 00:59

i always want u too
bloodlineofhate 00:59

(a) Chat excerpt. Original source [4]

Hey sexy.
ich_bin_der_eggman_67

How are you?
ich_bin_der_eggman_67

hi belvedere

:*
ich_bin_der_eggman_67

What have you been up to?
ich_bin_der_eggman_67 00:02

was sick belvedere 00:04

Awe no.
ich_bin_der_eggman_67 00:04

I am sorry :(
ich_bin_der_eggman_67 00:04

Do you want me to come make you feel better?
ich_bin_der_eggman_67 00:04

my aunti thinks is just cuz i miss my folks belvedere 00:04

lol belvedere 00:04

it wuld be nice but not rigte now aunti is home belvedere 00:05

Awe well that could be.
ich_bin_der_eggman_67 00:05

Yeah I would like to come see you ;)
ich_bin_der_eggman_67 00:05

(b) Chat excerpt. Original source [5]

Figure 6: Excerpts from full-length grooming chats with predictions by $S_{\text{BERT-large}}$ (for the messages in the respective excerpt only).