Implications of Using Internet Sting Corpora to Approximate Underage Victims

Tatiana Ringenberg, Kathryn Seigfried-Spellar, and **Julia Rayz**
Purdue Polytechnic
Purdue University
{tringenb, kspellar, jtaylo1} @purdue.edu

Abstract

Law enforcement officers (LEOs) and the justice system employ NLP models for classifying and triaging child exploitation cases due to the textual communications between predators and victims. The usefulness of these systems depend on the quality of data that can be used for training. Data in the domain are scarce, sensitive, and emotionally taxing for annotators. NLP researchers approximate victimization conversations using transcripts from internet stings performed by either vigilantes or LEOs, with an implicit assumption that vigilante or LEO conversations represent the victimization process. Psychology research, however, states that underage victim chats differ from internet stings in goal and modus operandi. We present a methodology and observations from annotating a corpus of victim, vigilante, and LEO conversations with convicted predators with the goal of comparing these chats. The corpus is annotated for stages and tactics of the victimization process described within psychology research. As predicted by psychological research, we found significant differences in the three classes of chats that are usually not taken into account in chat classification.

1 Introduction

Child exploitation crimes have expanded over the years to include a wide array of concerns including sextortion (Kopeckỳ, 2017), sex trafficking (Diaz and Panangadan, 2020), sexual solicitation (Briggs et al., 2011), and deep fakes (Albahar and Almalki, 2019). NLP research in understanding these chats is crucial because law enforcement agencies have become overwhelmed with online cases; automated systems are needed in order to sift through the available textual data and transcripts to improve case triage through identification of criminal activity (Inches and Crestani, 2012). In the past, automatic systems have been developed for differentiating predators from non-predators (Misra et al., 2019; Pendar, 2007), predicting level of risk throughout a conversation (Ringenberg et al., 2019), and flagging predatory conversations (Kim et al., 2020; McGhee et al., 2011; Zuo et al., 2018).

While the potential of NLP in the child exploitation domain is substantial, the corpora used to train these models is not always adequate. The data used to train algorithms rely on internet sting operations between an adult predator and a law enforcement officer (LEO) (DeHart et al., 2017), or adult vigilante (Black et al., 2015), impersonating a minor. Psychology research, which serves as the foundation for theory in the child exploitation domain, suggests internet sting operations progress differently than traditional conversations in which an adult seeks to victimize a child (DeHart et al., 2017; Briggs et al., 2011; Bergen et al., 2013; Mitchell et al., 2005). As a result, further research is needed to assess the ways in which these corpora differ and the potential impact the differences have on the resulting models.

We investigate to what extent internet sting operations may be assumed to accurately approximate child victimization. We focus on differences between sting participants and underage victims with respect to the victimization process known as online child grooming: the process an adult uses to gain the trust of a minor for the purpose of sexual fulfillment either online or in a physical meeting (O'Connell, 2003).

According to Stede and Huang (2012), "the most convincing evidence for the value of an annotation task remains to be its direct contribution to the success of one or more NLP applications" (p.92). This work highlights the potential negative impact of untested assumptions related to corpus composition in this domain. We also provide insights into the annotation process which may help others as-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3645–3656
August 1–6, 2021. ©2021 Association for Computational Linguistics

sess corpora within child exploitation. Finally we plan to release our code book, coinciding with this paper, as an aid to future work.

2 Online Child Victimization as a Process

The victimization process by which an adult entices a minor is referred to as grooming (O'Connell, 2003), and has been studied within the online (Kloess et al., 2019; O'Connell, 2003) and physical (Colton et al., 2010; Jackson et al., 2015) contexts for decades (Kaufman et al., 1993; Lang and Frenzel, 1988; Leclerc and Tremblay, 2007). We seek to understand the impact to NLP research. Thus, we limited the scope of our investigation to the textual transcripts produced from the online grooming process as originally defined by O'Connell (2003).

2.1 Characteristics of Online Grooming

Online grooming is the process an adult uses, on the Internet, to entice a minor into a sexual scenario (O'Connell, 2003). The process consists of six transient stages which include friendship forming, relationship forming, exclusivity, risk assessment, sexual, and meeting (O'Connell, 2003; Gupta et al., 2012). For a discussion of the grooming stages, see O'Connell (2003); Whittle et al. (2013). The process of grooming is not uniform; it ebbs and flows (Gillespie, 2002) based on the goals of the predator (Beauregard et al., 2012; Briggs et al., 2011; Kloess et al., 2017) and feedback from the victim (Wortley et al., 2019).

Common tactics used during online grooming include bragging (Aitken et al., 2018), compliments (Kloess et al., 2019), fantasy enactment (Kloess et al., 2019; Malesky Jr, 2007), coercion (Kloess et al., 2019; Villacampa and Gómez, 2017), repetition (Kloess et al., 2017), and expression of vulnerability (Barber and Bettez, 2014). For a thorough treatment on online grooming tactics, see Barber and Bettez (2014); Kloess et al. (2019). Predators use grooming tactics in order to further a goal or progress the relationship with the minor (Beauregard et al., 2012; Briggs et al., 2011; Kloess et al., 2017; Wortley et al., 2019).

2.2 Corpora in Online Grooming

To understand how datasets and corpora are built within the child exploitation domain, we performed a systematic review of research on child exploitation from January 2000 to March 2020. We searched for peer-reviewed journal articles in four databases: Medline, PubMed, PsychInfo, and ERIC. In each database, we searched for the terms *online sexual grooming*, *online sexual solicitation*, *child sexual abuse*, and *child molestation*. Results included applied NLP papers and psychology papers. Based on titles and abstracts, we selected papers which met the following criteria: peer-reviewed journal article, published in English, contained adult groomer and underage victims, clear online child grooming tactics component, and articles with empirical focus. Articles which were literature reviews or meta-analyses of grooming were removed.

We identified a total of 32 articles. We analyzed the type of participants and the location of the participants in the sample. Ten papers used vigilante cases, 16 used underage victim cases, one used interactions with LEO, and five contained a mixture of participants. These datasets are problematic because they may not represent the phenomenon being studied (Bowen, 2008; Hovy and Lavid, 2010). A representative corpus would ideally include the space of all participant types and behaviors (Bowen, 2008). However, transcripts from these three groups may differ in length (Briggs et al., 2011; Mitchell et al., 2005), motivation (Briggs et al., 2011; Williams et al., 2013), and modus operandi (Briggs et al., 2011; DeHart et al., 2017; Williams et al., 2013). Additionally, the studies were done on participants from a wide range of geographical jurisdictions including the United States (Black et al., 2015), the United Kingdom (Whittle et al., 2014), Sweden (Shannon, 2008), Spain (de Santisteban et al., 2018), Israel (Katz, 2013), or multiple countries (Quayle et al., 2014). This is problematic due to differences between countries with respect to the age of consent, legal ramifications, and social constructs surrounding child sexual abuse.

Online grooming corpora are generally used by ML researchers to identify specific grooming stages (Cano et al., 2014; Gupta et al., 2012) and classify conversations as predatory (Bogdanova et al., 2012; McGhee et al., 2011). Methods for identifying grooming tactics have included rule-based systems (McGhee et al., 2011), naïve bayes (Bogdanova et al., 2012), support vector machines (Gunawan et al., 2016), and more recently neural networks (Ebrahimi et al., 2016). Some of the common features used for classifying grooming conversations and stages include sentiment polarity

(Bogdanova et al., 2012; Cano et al., 2014), psycholinguistic categories (Cano et al., 2014; Gupta et al., 2012), and n-grams (Bogdanova et al., 2012).

2.3 Underage victims and internet stings: Could they differ?

Internet sting datasets do not represent the underage victimization process (Bergen et al., 2013; Briggs et al., 2011; Chiang and Grant, 2019; DeHart et al., 2017; Gijn-Grosvenor and Lamb, 2016; Mitchell et al., 2005; Schneevogt et al., 2018; Winters et al., 2017) as: 1) the type of predator caught in internet stings versus real victimization cases is different (Briggs et al., 2011; Mitchell et al., 2005); and 2) there are differences between the real vulnerabilities and reactions of at-risk minors versus trained vigilantes and LEOs with a goal of gathering evidence (Briggs et al., 2011; Chiang and Grant, 2019; DeHart et al., 2017; Gijn-Grosvenor and Lamb, 2016; Schneevogt et al., 2018; Williams et al., 2013). The first is impactful for NLP research because it affects the topics the predators pursue and the progress of stages. For instance, looking for sexually-charged language would likely be more effective in internet-stings than in underage victim chats (Briggs et al., 2011). The second is impactful because the data used to train algorithms is not representative of successful grooming techniques, especially if predators caught during stings were not familiar with grooming tactics for luring minors. These factors affect how openness and directness of communication is used by law enforcement and vigilantes in comparison to underage victims (Briggs et al., 2011; Williams et al., 2013).

Conversely, DeHart et al. (2017) found many internet-based predators requested meetings with victims. While more research is needed, empirical evidence shows the predators in internet stings may be different enough to impact how the online grooming process manifests (Briggs et al., 2011; Mitchell et al., 2005).

Psychologists also posited several ways in which the grooming process may differ in terms of modus operandi. Briggs et al. (2011); DeHart et al. (2017); Williams et al. (2013) claimed LEO, or vigilantes, are more likely than victims to be open to online sexual behavior and requests to meet (Briggs et al., 2011; DeHart et al., 2017). Further, authors posit LEOs or vigilantes differ with respect to search criteria (DeHart et al., 2017), reaction to sexual comments (Briggs et al., 2011; Williams et al., 2013), specialized training (Briggs et al., 2011), overt explicitness of profiles (Briggs et al., 2011), and coercion (Chiang and Grant, 2019; Schneevogt et al., 2018). Finally, vigilantes and LEO are limited in what they can say to secure an arrest, as they are discouraged from initiating sexual conversation, contact, and arrangements of meetings (Gijn-Grosvenor and Lamb, 2016). As a result, LEOs and vigilantes may attempt to nudge conversations in this direction (Williams et al., 2013).

Finally, Briggs et al. (2011) highlighted the importance of the difference in goals between underage victims, vigilantes, and LEOs. Briggs et al. (2011) noted vigilantes and LEOs have a goal of collecting evidence and securing a quick arrest. This results in shorter chats with faster progressions and more pointed language than victim chats (Briggs et al., 2011). In LEO and vigilante chats, there was eagerness which is not reflective of either the distrust or language used by at-risk teens (Briggs et al., 2011).

3 Methodology

The annotation protocol described in the following sections follows the annotation pipeline proposed by Hovy and Lavid (2010). Based on research presented above, we hypothesize the greatest actionable differences occur within the grooming stages and grooming tactics actualized by underage victim, vigilante and LEO chats.

3.1 Corpus Composition: Data Subjects

A corpus is considered to be representative once it is *saturated* (Bowen, 2008), such that all relevant aspects of the phenomena are covered (Bowen, 2008; Hovy and Lavid, 2010) to the point of repetition or redundancy (Bowen, 2008). Per our assessment of online grooming literature in the 2000s, datasets in the predator domain consist of underage victims, LEOs, and online vigilantes. As data in the domain is difficult and time-consuming to acquire, we are unable to construct a corpus which would satisfy the saturation metric. However, we are able to construct a small corpus which represents all three participant types within the domain. We constructed a corpus of the three groups to ensure we account for all participants. Our corpus consists of 60 chat transcripts representing an equal number of participants from the three groups: 20 vigilantes, 20 underage victims, and 20 LEOs.

3.2 Data Acquisition and Cleaning

Our data collection and data use practices adhere to our Institutional Review Board (IRB) protocol. Due to the sensitive nature of the data, we submitted a full protocol review, which was approved. All texts in the corpus had all images and identifying information removed.

Underage victim and LEO transcripts were obtained from local, state, and federal agencies within the United States. While previous studies consisted of transcripts from multiple countries (Bergen et al., 2013; Quayle et al., 2014), we focused on transcripts from a single country in order to minimize confounding variables which could influence how participants interact.

The vigilante conversations are publicly available through pervertedjustice.com, an organization of adult vigilantes, trained to pose as minors online. The vigilantes speak with predators and work with LEOs to secure evidence for convictions. Oversight of these individuals is minimal (Williams et al., 2013) and there is little information on their website concerning the content of training for the vigilantes. The chat transcripts are posted to the Perverted Justice website following conviction.

One limitation we identified was incompleteness of the data. Transcripts from LEOs rarely included the full interaction between participants. Additionally, vigilante conversations would occasionally reference a phone call which took place without making the call transcript available. We also found that some chats ended early when LEOs had enough evidence. Other chats took place partway through the interaction. Still others terminated when the LEO switched to a different chat service.

Our mitigation for incomplete transcripts was to request background information from the law enforcement agencies. In the case of Perverted Justice transcripts, case documents and transcript summaries are often linked to the chats on the site. This can be a helpful resource for identifying limitations of individual transcripts.

3.3 Constructing the Initial Code Book

Tactics in our code book were deductively selected from grooming stages and tactics which map to the limitations discussed in Section 2.3.

The consensus from Section 2.3 was the grooming process would be affected by participant goals and investigator tactics (Bergen et al., 2013; Williams et al., 2013). To capture these differences, we considered all stages outlined by O'Connell (2003). The stages included *friendship forming*, *relationship forming*, *exclusivity*, *sexual*, *risk assessment*, and *meeting*. Additionally, as we annotated the stages, we noticed a possible divide between the *sexual stage* and a *non-consensual sexual stage* in which the victim, vigilante, or LEO indicated discomfort or declined sexual advances.

We selected the grooming tactics to operationalize the limitations enumerated in Section 2.3. The tactics are summarized in Table 1.

Openness to sexual comments/behavior. The *bragging* and *personal compliment* tactics are linked to sexual discussions within grooming (Barber and Bettez, 2014). Given LEOs respond to sexual content in a more positive manner than at-risk teens, we posit responses related to *bragging* and *personal compliments* will differ. Roleplay is how some predators act out fantasy, thus making it applicable to assessing openness to sexual behavior (Kloess et al., 2019). Finally, predators often ask for jarring and explicit details about the victim's sexual past (Aitken et al., 2018) which an at-risk teen might find uncomfortable.

Discussion of Meetings. *Willingness* is an assessment of what a participant would consider doing (Barber and Bettez, 2014). Often, this is associated with discussions of the explicit activities a victim would agree to perform in person.

Coercion. *Coercion* has been associated with differences between internet stings and victim conversations in the past, though they often point to more overt forms of coercion (Chiang and Grant, 2019; Schneevogt et al., 2018). We split the *coercion* tactic into *coercion* and *sexual violence* to reflect this distinction.

Naïve and Young/Explicitness of Profile. Vulnerabilities are used by predators to evoke sympathy (Barber and Bettez, 2014). The expression of vulnerabilities by a non-predator may be a good indicator that a sting is taking place, as we hypothesize LEOs and vigilantes will over-emphasize childhood problems. Further, willingness to send images in chat to a stranger could also be an operationalization of naïvety to online dangers. We posit LEOs and vigilantes are more likely to send such pictures than real at-risk minors.

Initial Contact. While we could not find any grooming tactics related to initial contact within the literature, we posit differences in initial contact can be captured by the grooming stages.

Grooming Characteristics	Grooming Tactics
Openness to sexual comments/behavior	Roleplay Bragging Sexual History Personal Compliments
Discussion of Meetings	Willingness
Coercion	Coercion Sexual Violence
Naïve/Explicitness of Profile	Discussion of Images Vulnerabilities (Neg. Life Stories, Neg. Physical Attributes)
Initial Contact	Grooming Stages
Arrest Goal	Age Difference Reverse Power
Speed/Duration	Media Progression (Phone calls, Video Chat)

Table 1: Mapping of grooming characteristics to grooming tactics in this paper.

Arrest Goal. We focused on what an LEO would need to show that a law was broken. Stating the *age difference* would indicate the predator knew the victim was under 18. *Reverse power* is giving the control of the situation to the other participant (Barber and Bettez, 2014). We hypothesize there may be differences with respect to *reverse power* because the vigilantes and LEOs are unable to initiate certain discussions (Gijn-Grosvenor and Lamb, 2016). *Reverse power* may be a way to influence the discussion without explicitly initiating.

Speed/Duration. In typical grooming conversations, participants often switch chatting services multiple times (Quayle et al., 2014). If conversations are shorter and faster in internet stings, we posit there will be fewer forms of communication. In addition to definitions for each of the tactics and stages from literature, code books often include examples for each tactic (Bada et al., 2012; Kingsbury et al., 2002; Stoyanov and Cardie, 2008) which help to clarify the tactic without over-specifying (Hovy and Lavid, 2010). In the code book we included an initial set of examples for each tactic from the psychology literature (2.1).

3.4 Annotator Selection

Grooming stages and tactics are difficult to annotate, even for those experienced within the domain (Gillespie, 2002). Recommendations of whether to use domain experts or novices differs within NLP literature (Hovy and Lavid, 2010). We used two annotators: the first annotator is an expert within the child exploitation domain while the second annotator was an undergraduate with low levels of experience. Due to the multidisciplinary nature of this domain, we decided to use a non-expert for the second annotator to ensure our corpus was accessible to researchers with multiple backgrounds.

3.5 Annotator Training

Hovy and Lavid (2010) recommends providing a reasonable amount of training in order to familiarize the annotators with the task without over-specifying the task. To accomplish this, the first annotator gave the novice second annotator a short overview of the child exploitation domain to help bridge the gap between skill sets to improve alignment (Bayerl and Paul, 2011).

The first annotator also provided the second annotator with the initial code book to review. The second annotator was able to ask clarifying questions about the grooming stages and tactics in the code book.

Following familiarization with the codebook, both annotators annotated a series of three training chats. Since grooming participants do not use every single grooming stage or tactic in chats, but rather a subset that fits their goals (Beauregard et al., 2012; Briggs et al., 2011; Kloess et al., 2017), annotating three chats ensured exposure of the second annotator to all of the stages and tactics in the code book.

Simple agreement was used to calculate a baseline of agreement on which to measure improvement following training; the target agreement for this study was 80% (Krippendorff, 2011). Per best practices, the annotators met to discuss the training round, answer questions from the second annotator, and resolve inter-annotator disagreements (Hovy and Lavid, 2010).

3.6 Annotation Procedures

3.6.1 Coding Rounds

We initially had six grooming stages and 14 grooming tactics to annotate. Due to the expected complexity of these tactics and stages, we originally split the annotation task into three rounds. Following the training round we re-split the task into four rounds to reduce annotator fatigue in an annotation session (Bayerl et al., 2003).

One recommendation from this is to keep the grain-size of stages or tactics consistent within a single annotation round. In the first round of annotation we had a mixture of stages and tactics. However, this made annotation difficult for two reasons. The first reason was that grooming stages encompassed grooming tactics. For instance, within the *sexual stage* a predator might use *willingness* as a means to gauge interest and *roleplay* to normalize sexual content. We found the difference in granularity between grooming stages and tactics made it difficult to context switch within the same round. The second reason was that grooming stages were complex and impermanent (Gillespie, 2002). There were multiple examples and rules to remember with each grooming stage and changes between grooming stage were often quick and overlapped, making the annotation task complicated. Keeping the number of tactics low made it easier to focus on the stages.

3.6.2 Code Span Identification

The transient nature of grooming makes the beginning and end of a grooming instance difficult to identify (Gillespie, 2002). Known as *code span*, challenges related to identification of the boundaries of an annotation are known issues within annotation literature (Bada et al., 2012; Stoyanov and Cardie, 2008). To mitigate code span issues, we allow partial membership to a tactic or stage to facilitate discussion and identify span issues.

For this study, we defined *partial membership* as (i) non-substantive statements made in reference to a tactic or (ii) lines which do not meet the criteria for a tactic but result in or from the tactic. Full membership was defined as a line which was representative of a tactic. For instance, if a victim indicates their parents will be gone for the weekend and the predator responds by asking if they would like to meet, we would annotate the statements about the meeting as having full membership to the *meeting stage*. We would also annotate the victim's line about their parents as partial membership because the message precipitated the meeting discussion.

Based on annotating the corpora, we did not find a noticeable difference in the code span of topics between victim, vigilante, and LEO transcripts. However, we did notice that for longer spans of sexual tactics such as the *sexual stage*, *sexual history*, and *roleplay*, the intensity and graphic content of the messages would increase. We posit this was due to the predators attempting to gradually sexualize the conversation. Given that LEO and vigilantes respond in a more positive manner that real victims (Briggs et al., 2011; DeHart et al., 2017; Williams et al., 2013), predators may use more graphic and descriptive language when talking to vigilantes and LEOs than real victims. Additionally, we noticed real victims would end uncomfortable conversations more quickly and with more firm language (e.g., "no" versus "idk"). Future research should examine the change in graphical descriptions and responses to them over time in transcripts.

3.6.3 Annotator Protocol and Process

We developed the following guidelines for the annotation process:

- Review the code book before each annotation session.
- Annotation questions should be noted at the time, and discussed between rounds.
- Questions should be directed towards the first annotator and not external works or researchers. Hovy and Lavid (2010) emphasized the issues which could arise from inexperienced annotators not knowing the correct resource to ask for assistance. We mitigate this by having the first annotator, with domain experience, be the sole source of training.
- Annotators should label each round independently; annotations do not depend on previous rounds. We added this rule following the training session because the non-expert annotator began to second-guess tactics and annotations they felt should be related to one another.
- Coding session duration should be limited to reduce fatigue (Bayerl et al., 2003). We limited coding sessions to a maximum of three rounds in an annotation session.
- Each line may be annotated for multiple tactics within the coding round. While limiting annotations to one per tactic per round is desirable (Bada et al., 2012), the complex and transient nature of grooming results in multiple tactics and stages manifesting at the same time (Gillespie, 2002).

Using these guidelines and the finalized code book, both annotators annotated six chat transcripts, which represented 10% of the total number of transcripts. Simple agreement was calculated and found to be greater than the recommended 80% for annotation tasks (Krippendorff, 2011). For the

Group	Total Lines	*M*	*SD*
Vigilante	28505	1425.25	1111.96
Victim	10469	523.45	742.74
LEO	5140	257.00	302.56

Table 2: Summary of corpus composition.

simple agreement calculation, we used the total agreed upon annotations out of the total annotations. For this study, we only took into account full membership annotations as partial membership annotations were created to facilitate discussion. Future work, which incorporates the partial membership annotations into analysis, would require additional agreement measures.

Differences following the agreement calculation were each discussed and resolved through verbal agreement.

Finally, the first annotator annotated the remaining 51 chat conversations using the finalized code book and above guidelines. Future work should use two or greater annotators for the duration of annotation (Bayerl and Paul, 2011). While using a single annotator for the remainder of the annotation is a limitation of this study, we sought to mitigate the limitation by constructing and testing the finalized code book using both annotators.

3.6.4 Comparing Victims, Vigilante, and LEOs

Quantitative Analysis

To compare chat length, we used a one way analysis of variance between the three groups. Levene's F Test indicated unequal variance ($F = 10.82, p < .001$). As a result, the Welch Test was performed and showed a significant effect of participant type on length of chat, *Welch's* $F(2, 57) = 10.587, p < .001$, $\omega = .52$. We performed Post-hoc analysis using the Games-Howell test due to the unequal variance. The Games-Howell test indicated the number of lines for a vigilante chat were significantly different than the number of lines for LEO ($p < .001$) and underage victims ($p = .013$).

. A χ^2 test was conducted to examine the presence or lack of each tactic within chats in the three groups. While the majority of chats contained all stages, future work should assess differences in presence and sequencing of grooming stages as well.

When assumptions of a χ^2 test could not be met, a Fisher's Exact test was conducted. Through χ^2 tests, we found *sexual history* ($\chi^2 = 10.40, df = 2, p = .006$), *willingness* ($\chi^2 = 14.55, df = 2, p = .001$), *discussion of phone calls* ($\chi^2 = 10.15, df = 2, p = .006$), and acknowledgement of *age difference* ($\chi^2 = 22.67, df = 2, p < .001$) were significantly different between participant groups. Through Fisher's Exact tests, we found the use of *personal compliments* ($p = 0.011$) and *reverse power* ($p = 0.03$) were significantly different between participant groups. There were no significant differences between the three groups with respect to *coercion*, *discussion of images*, *bragging*,*discussion of video chatting*, *negative physical traits*, *negative life stories*, *roleplay*, or *sexual violence*.

	Vigilante	Victim	LE
Sexual History	95.0%	50.0%	75.0%
Willingness	85.0%	55.0%	25.0%
Phone Calls	75.0%	55.0%	25.0%
Age Difference	95.0%	35.0%	90.0%
Compliments	100.0%	65.0%	83.3%
Reverse Power	100.0%	75%	70%

Table 3: Occurrence of grooming tactics in vigilante, victim, and LEO conversations.

The results of the χ^2 and Fisher's Exact tests are summarized in Table 3.

Within the three groups, grooming tactics were used the most in vigilante conversations. This was consistent for all tactics in Table 3. Most conversations with vigilantes included the use of all of the tactics in Table 3. *Willingness*, *discussion of phone calls*, and *reverse power* were used the least in LEO chats. *Sexual history*, *age difference*, and *compliments* were used the least in victim conversations.

LEO conversations tended to be short and often included a shift to talking in another app. Additionally, LEO responses to direct questions about meeting and sexual activities tended to be vague in comparison to vigilantes and real victims. The vagueness of the LEO responses may have resulted in the predator adjusting tactics.

We posit age difference and sexual history may be used less frequently in victim conversations because the victims often already knew the predator. The sexual history of the victim may already be known by the predator. This is likely the case with age difference as well. The victims and predators often appeared aware of the age gap but would not discuss it. In vigilante and LEO conversations, the conversations were with strangers. The participants would often give their age, sex, and location towards the beginning of the chat.

We originally posited LEO and vigilante conversations would be similar. Both vigilante and LEO groups consist of adults posing as children. Additionally, both groups receive some level of dedicated training on identifying predators. However, from Table 3 we see the groups differ in the tactics used. Some of these differences may be the result of the motivation of the LEO versus the vigilante and some may be a result of the differences in training. In the next section, we will discuss our qualitative observations which may contribute to differences in both how and how often the grooming tactics were used.

Qualitative Analysis

In addition to the analysis of chat length and tactic usage, we also qualitatively noted, and aggregated, annotator observations on language and conversational differences between the three groups. The differences we found between real victims and internet stings were the result of the LEOs and vigilantes attempting to make the predator state intentions explicitly and avoid sexual situations or meetings.

In many cases, individual lines taken out of context would appear innocuous. For instance, it is common for a predator to ask for images of a victim to get to know them and determine level of attraction. However, based on this exchange alone, it would not be possible to determine whether or not one of the participants were a predator. In these cases, LEOs and vigilantes would often ask clarifying questions to determine whether the request was for sexual or non-sexual pictures. At times, the LEOs or vigilantes would go as far as to describe the clothing they were wearing, such as making references to their pajamas, when the predator asked for pictures. We observed this as a priming technique to prompt the user to ask for more graphic images. Additionally, we found real victims were more likely to provide images while vigilantes and LEOs tended to ask clarifying questions around image requests.

Furthermore, it is also common for a predator to request images when attempting to determine if the other participant is an LEO instead of a minor (Kloess et al., 2019). In some cases, the predator will demand pictures taken immediately to ensure the participant is the person in the image.

From the chats we annotated, this appeared to occur more frequently in predator and vigilante chats than in underage victim chats. We hypothesize that the LEOs or vigilantes were being forward in a manner that sparked suspicion in the predator. Additionally, this was often triggered by another tactic in which the officer or vigilante would say they lied about their age in the profile and were actually younger. Statements like this did not occur in victim chats and seemed to cause suspicion and trigger risk assessment questions from the predator.

We also found direct and indirect communication styles used by the predator affected the three participant groups differently. Predators would often use expressions of vagueness about intentions to avoid explicitly stating an interest in sex. Further, we saw examples in which predators would not refer to sexual body parts or sexual acts directly, but instead would use euphemisms.

LEO and vigilantes would respond to vague responses about meetings, sexual intentions, and euphemisms by repetitively asking the predator what they meant or what they would do together. Additionally, the LEOs and vigilantes would act naïve and ask the predator to explain obvious references to body parts or sexual innuendos. We did not see this within underage victim chats. In real victim chats, the predators appeared to do the majority of the prompting related to sexual topics. We did not see much evidence of victims asking predators what they would do when they met. However, this was present in most vigilante and officer chats where meetings were discussed.

The greatest disparities between vigilantes and LEOs were within the *willingness* and *phone call* tactics. *willingness* to assess whether or not the vigilantes would engage in specific sexual activities during a physical meeting. In some instances, predators would also use *willingness* to get the other participant to agree to follow their instructions. We posit the difference in occurrence of the *willingness* tactic may be due to the responses given. Vigilantes and victims appear to respond more positively to *willingness* questions whereas LEOs tend to respond more vaguely. This may lead the predator to adjust tactics when speaking to LEOs.

Finally, LEOs and vigilantes used contrived situations to avoid undesired interactions or unplanned meetings with a predator. Such interactions included roleplaying, sending sexual images, or watching sexual activity on a webcam. Examples of avoidances used for undesirable interactions included an angry parent, a parent in the room, homework, a broken phone or camera, a broken internet connection, tiredness, or plans with friends. Avoidances of meetings were generally framed around other events preventing the meeting. For instance, if a predator wanted to meet the vigilante or LEO but the participant was not ready, the vigilante or LEO would claim to have family plans that were unavoidable. Avoidances were rarely used by underage victims.

3.7 Iterative Changes to Corpus

Neutering is a common strategy to handle disagreement and combine overlapping tactics, and refers to collapsing two or more tactics together in a code book (Hovy and Lavid, 2010). During the annotation process, we chose to neuter two sets of tactics: sexual stage and non-sexual stage; and relationship forming and exclusivity.

As described in 3.5, we treated the *sexual stage* and the *non-consensual sexual stages* as separate stages during annotator training. Following training, we neutered the stages into a single stage as the annotators often could not agree on what constituted an implied denial of an advance. Future research should investigate the possible presence of a *non-consensual sexual stage* where the victim implicitly or explicitly scorns advances.

Given the similarity between *relationship forming* and *exclusivity*, we also neutered these stages into a single stage. *Relationship forming* and *exclusivity* revolve around the construction of foundational trust between the predator and the victim (O'Connell, 2003). While relationship forming can be thought of as the day-to-day interactions, exclusivity revolves around language to intensify the relationship and isolate the victim from their support system (O'Connell, 2003).

In addition to neutering tactics within the code book, we added examples for each tactic. The original stages and tactics in the literature were designed from the perspective of the predator (Barber and Bettez, 2014), kloess2017qualitative, O'Connell. We annotated each of the tactics for all participants. We found the way LEOs, vigilantes, and underage victims used and responded to tactics differed from the predator. For instance, the predators used references to *age difference* to determine the level of comfort of the victim while LEOs used *age difference* to ensure the predator explicitly acknowledged the illegality of the solicitation. Further, underage victims would sometimes reference the age gap as a negative trait of themselves, almost as an insecurity. Having examples of uses of the tactics and stages by multiple types of participants helped the second annotator to generalize the tactics to all participants and not just predators.

4 Conclusion

NLP has contributed to research in the child exploitation domain by developing automated systems for detection of predators (Kim et al., 2020; McGhee et al., 2011; Zuo et al., 2018) and participants (Pendar, 2007). However, the corpora used for training the models are not representative of the criminals or victims involved in actual child exploitation cases (Bergen et al., 2013; Briggs et al., 2011; Chiang and Grant, 2019; DeHart et al., 2017; Gijn-Grosvenor and Lamb, 2016; Mitchell et al., 2005; Schneevogt et al., 2018; Winters et al., 2017). We offered an overview of the problems within the NLP corpora in the domain. We also discussed the impact these problems have on representing the online grooming process. Finally, we provided our methodology and recommendations from annotating a corpus of underage victim, vigilante, and LEO conversations and showed that there are statistical differences between the three groups. While NLP research within the child exploitation domain appears to be expanding, there is a need to ensure that corpora are designed and annotated in such a way that it contributes beneficial solutions.

Acknowledgments

We would like to thank the reviewers for their feedback and recommendations. Additionally, we would like to thank Lin Sun Fa for his assistance as the second annotator for this project.

References

Susan Aitken, Danielle Gaskell, and Alan Hodkinson. 2018. Online sexual grooming: exploratory comparison of themes arising from male offenders' communications with male victims compared to female victims. *Deviant Behavior*, 39(9):1170–1190.

Marwan Albahar and Jameel Almalki. 2019. Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22):3242–3250.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and et al. 2012. Concept annotation in the craft corpus. *BMC Bioinformatics*, 13(1):161.

Connie Barber and Silvia Bettez. 2014. Deconstructing the online grooming of youth: Toward improved information systems for detection of online sexual predators.

Petra S Bayerl, Harald Lüngen, Ulrike Gut, and Karsten I Paul. 2003. Methodology for reliable schema development and evaluation of manual annotations. In *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003)*.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Eric Beauregard, Benoit Leclerc, and Patrick Lussier. 2012. Decision making in the crime commission process: Comparing rapists, child molesters, and victim-crossover sex offenders. *Criminal Justice and Behavior*, 39(10):1275–1295.

Emilia Bergen, Jan Antfolk, Patrick Jern, Katarina Alanko, and Pekka Santtila. 2013. Adults' sexual interest in children and adolescents online: A quasi-experimental study. *International Journal of Cyber Criminology*, 7(2).

Pamela J Black, Melissa Wollis, Michael Woodworth, and Jeffrey T Hancock. 2015. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child abuse & neglect*, 44:140–149.

Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. 2012. On the impact of sentiment and emotion based features in detecting online sexual predators. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 110–118.

Glenn A Bowen. 2008. Naturalistic inquiry and the saturation concept: a research note. *Qualitative research*, 8(1):137–152.

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):72–91.

Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani. 2014. Detecting child grooming behaviour patterns on social media. In *International conference on social informatics*, pages 412–427. Springer.

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.

Matthew Colton, Susan Roberts, and Maurice Vanstone. 2010. Sexual abuse by men who work with children. *Journal of child sexual abuse*, 19(3):345–364.

Dana DeHart, Gregg Dwyer, Michael C Seto, Robert Moran, Elizabeth Letourneau, and Donna Schwarz-Watts. 2017. Internet sexual solicitation of children: a proposed typology of offenders based on their chats, e-mails, and social network posts. *Journal of Sexual Aggression*, 23(1):77–89.

Maria Diaz and Anand Panangadan. 2020. Natural language-based integration of online review datasets for identification of sex trafficking businesses. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 259–264. IEEE.

Mohammadreza Ebrahimi, Ching Y Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, 18:33–49.

Evianne L. van Gijn-Grosvenor and Michael E. Lamb. 2016. Behavioural differences between online sexual groomers approaching boys and girls. *Journal of Child Sexual Abuse*, 25(5):577–596.

Alisdair A Gillespie. 2002. Child protection on the internet-challenges for criminal law. *Child & Fam. LQ*, 14:411.

Fergyanto E Gunawan, Livia Ashianti, Sevenpri Candra, and Benfano Soewito. 2016. Detecting online child grooming conversation. In *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, pages 1–6. IEEE.

Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. 2012. Characterizing pedophile conversations on the internet using online grooming.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. page 25.

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.

Sharon Jackson, Elinor Newall, and Kathryn Backett-Milburn. 2015. Children's narratives of sexual abuse. *Child & Family Social Work*, 20(3):322–332.

Carmit Katz. 2013. Internet-related child sexual abuse: What children tell us in their testimonies. *Children and Youth Services Review*, 35(9):1536–1542.

Keith L Kaufman, Daniel R Hilliker, Patty Lathrop, and Eric L Daleiden. 1993. Assessing child sexual offenders' modus operandi: Accuracy in self-reported use of threats and coercion. *Annals of Sex Research*, 6(3):213–229.

Jinhwa Kim, Yoon Jo Kim, Mitra Behzadi, and Ian G Harris. 2020. Analysis of online conversations to detect cyberpredators using recurrent neural networks. In *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*, pages 15–20.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the penn treebank. In *Human Language Technology Conference*.

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):73–96.

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):563–591.

Kamil Kopeckỳ. 2017. Online blackmail of czech children focused on so-called "sextortion"(analysis of culprit and victim behaviors). *Telematics and Informatics*, 34(1):11–19.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Reuben A Lang and Roy R Frenzel. 1988. How sex offenders lure children. *Annals of Sex Research*, 1(2):303–317.

Benoit Leclerc and Pierre Tremblay. 2007. Strategic behavior in adolescent sexual offenses against children: Linking modus operandi to sexual behaviors. *Sexual Abuse: A Journal of Research and Treatment*, 19(1):23–41.

L Alvin Malesky Jr. 2007. Predatory online behavior: Modus operandi of convicted sex offenders in identifying potential victims and contacting minors over the internet. *Journal of child sexual abuse*, 16(2):23–32.

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*, 15(3):103–122.

Kanishka Misra, Hemanth Devarapalli, Tatiana R Ringenberg, and Julia Taylor Rayz. 2019. Authorship analysis of online predatory conversations using character level convolution neural networks. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 623–628. IEEE.

Kimberly J Mitchell, Janis Wolak, and David Finkelhor. 2005. Police posing as juveniles online to catch sex offenders: Is it working? *Sexual Abuse: A Journal of Research and Treatment*, 17(3):241–267.

Rachel O'Connell. 2003. A typology of child cybersexploitation and online grooming practices. *Cyberspace Research Unit, University of Central Lancashire*.

Nick Pendar. 2007. Toward spotting the pedophile telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*, pages 235–241. IEEE.

Ethel Quayle, Silvia Allegro, Linda Hutton, Michael Sheath, and Lars Lööf. 2014. Rapid skill acquisition and online sexual grooming of children. *Computers in Human Behavior*, 39:368–375.

Tatiana R Ringenberg, Kanishka Misra, and Julia Taylor Rayz. 2019. Not so cute but fuzzy: Estimating risk of sexual predation in online conversations. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2946–2951. IEEE.

Patricia de Santisteban, Joana del Hoyo, Miguel Ángel Alcázar-Córcoles, and Manuel Gámez-Guadix. 2018. Progression, maintenance, and feedback of online child sexual grooming: A qualitative analysis of online predators. *Child Abuse Neglect*, 80:203–215.

Daniela Schneevogt, Emily Chiang, and Tim 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*, 5(1):97–102.

David Shannon. 2008. Online sexual grooming in sweden—online and offline sex offences against children as described in swedish police data. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 9(2):160–180.

Manfred Stede and Chu-Ren Huang. 2012. Interoperability and reusability: the science of annotation. *Language Resources and Evaluation*, 46(1):91–94.

Veselin Stoyanov and Claire Cardie. 2008. Annotating topics of opinions. page 6.

Carolina Villacampa and Mª Jesus Gómez. 2017. Online child sexual grooming: Empirical findings on victimisation and perspectives on legal requirements. *International Review of Victimology*, 23(2):105–121.

Helen Whittle, Catherine Hamilton-Giachritsis, Anthony Beech, and Guy Collings. 2013. A review of online grooming: Characteristics and concerns. *Aggression and violent behavior*, 18(1):62–70.

Helen C Whittle, Catherine E Hamilton-Giachritsis, and Anthony R Beech. 2014. “under his spell”: Victims’ perspectives of being groomed online. *Social Sciences*, 3(3):404–426.

Rebecca Williams, Ian A Elliott, and Anthony R Beech. 2013. Identifying sexual grooming themes used by internet sex offenders. *Deviant behavior*, 34(2):135–152.

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):62–76.

Richard Wortley, Benoit Leclerc, Danielle M Reynald, and Stephen Smallbone. 2019. What deters child sex offenders? a comparison between completed and noncompleted offenses. *Journal of interpersonal violence*, 34(20):4303–4327.

Zheming Zuo, Jie Li, Philip Anderson, Longzhi Yang, and Nitin Naik. 2018. Grooming detection using fuzzy-rough feature selection and text classification. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE.