2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)
Bari, Italy. October 6-9, 2019

Not So Cute but Fuzzy: Estimating Risk of Sexual Predation in Online Conversations

Tatiana R. Ringenberg*, Kanishka Misra* and Julia Taylor Rayz
Department of Computer and Information Technology, Purdue University, USA

Abstract— **The sexual exploitation of minors is a known and persistent problem for law enforcement. Assistance in prioritizing cases of sexual exploitation of potentially risky conversations is crucial. While attempts to automatically triage conversations for the risk of sexual exploitation of minors have been attempted in the past, most computational models use features which are not representative of the grooming process that is used by investigators. Accurately annotating an offender corpus for use with machine learning algorithms is difficult because the stages of the grooming process feed into one another and are non-linear. In this paper we propose a method for labeling risk, tied to stages and themes of the grooming process, using fuzzy sets. We develop a neural network model that uses these fuzzy membership functions of each line in a chat as input and predict the risk of interaction.**

I. INTRODUCTION

The sexual exploitation of online youths is a known and persistent problem. The National Center for Missing and Exploited Children (NCMEC) received 10.2 million reports of suspected child exploitation in 2017 [1]. Researchers have suggested the exploitation of minors online is also likely under-reported, due in part to the offender encouraging the minor to keep the interaction a secret and the minor acquiescing to the request [2]. Furthermore, handling the cases we even know about has placed additional strain on law enforcement agencies[3], and any help with prioritizing high risk conversations is a step forward.

Since the early 2000s, researchers have devoted a considerable amount of time and resources towards understanding and modeling the online grooming process which occurs between adults and minors [4], [5]. Studies related to online grooming theory are focused on the techniques and themes used within the various stages of the process [4], [6], offender and victim characteristics [7], and offender identification [8], [9]. While qualitative analyses have identified many of the themes and patterns occurring within each of the grooming stages [4], [10], little research on offender identification uses this information. The majority of research on offender identification uses low-level features (n-grams, word categories, basic chat statistics) which have been found to be less effective than the use of high-level features directly related to grooming [11]. There is a growing need for research which focuses on the use of grooming-derived features to improve the identification of online solicitors of minors.

While the use of high-level features is ideal, expert annotation is required due to the complexity of the task [12]. Even for a researcher in online grooming, the task of identification of various aspects of the grooming process is difficult because the stages are neither linear nor perfectly discrete [4], [6]. From previous research, we know within the first 20% of a conversation, an offender may engage in several stages of the grooming process including friendship forming, risk assessment, and the sexual stage. We also know the progression through the grooming stages can be gradual both within and between stages[6]. For instance, within the sexual stage offenders will often start with innocuous questions related to a minor's previous romantic relationships but will progress to questions about sexual history, hypotheticals about future encounters, and at times graphic sexual descriptions of fantasy [10], [13]. Furthermore, identification of code span is a known problem within Natural Language Processing literature and is further compounded by the use of chat logs in which subsequent lines may or may not refer to the line prior [14], [15]. As the bounds between stages are not discrete and the code span of a given topic is problematic due to the nature of the conversation structure, fuzzy annotation of grooming-related content is ideal. Through the use of fuzzy annotation, researchers can model the gradual progression and oscillation of various aspects of the grooming process.

We address the gap in offender literature by creating fuzzy annotations for a small corpus of grooming conversations, where each message is mapped to three levels of risk associated with the grooming process, namely low, medium and high. The novelty of our approach is in modeling imprecise boundaries of annotations of risk in conversations of offenders and minors. This increases the speed of annotation, thus making it possible to process time sensitive chats that may be used for training or development set faster. A chat message can belong to multiple categories with varying degrees of membership. We employ fuzzy sets [16] to approximate the membership.

II. RELATED WORK

Identification of predatory intent has been studied from both a qualitative and computational perspective. Within qualitative research, researchers have identified the processes and themes offenders use to entrap minors [4], [5], [17]. From the computational perspective, researchers have worked on the identification of non-offenders from offenders [8], [9] along with identifying various aspects of the grooming stages [11].

O'Connell [4] developed the online grooming process which consisted of the following stages: *friendship forming,*

* First two authors contributed equally

978-1-7281-4569-3/19/$31.00 ©2019 IEEE

relationship forming, risk assessment, exclusivity, and sexual. The friendship forming stage consists of conversation which would generally be considered normal. Relationship forming is also non-sexual but includes more child-specific themes than the friendship forming stage. Risk assessment revolves around reducing the probability of detection. The exclusivity stage is the stage in which a bond with the child is established. Finally, the sexual stage is a complex amalgamation of topics and techniques an offender uses to normalize sexual discussion. Offenders normalize sexual discussion by starting with innocuous questions and ramping up sexual discussion to include hypotheticals, cyber sex, and at times images [4]. Gupta et al. [17] labeled the stages identified by O'Connell in 75 chat conversations to identify patterns in and between phases. The authors suggested the relationship forming stage is the most dominant stage. Additionally, the authors found the topic of meeting did not occur only at the end of the chat but rather occurred in multiple places throughout the chat [17].

Other studies have focused on the themes which occur within the chat conversations. Kloess et al. [13] examined a set of five cases consisting of 29 transcripts to find patterns related to the modus operandi of the offender. The authors found themes related to initiating online sexual activity, pursuing sexual information, fantasy rehearsal, and discussion of physical meeting [13]. Barber and Bettez [10] also annotated a series of conversations for grooming themes and five main themes: assessment, enticements, cyberexploitation, control, and self-preservation.

Aside from grooming stages and themes, research on offender strategies has provided clues as to the pacing and progress of the offender process. As far back as the 1970s and 1980s, authors discussed the concept of progression within the grooming process. Peters [18] described the movement of offenders towards sexual activity with a minor as the result of the affection-seeking of the child. Additionally, Lang and Frenzel [19] identified progressive horseplay as a means to move from innocuous childrens' games to seemingly accidental touching which finally results in sexual contact. The authors also found the progression of the sexual activity was not always gradual - the offender would intensify efforts when it appeared the child was either confused or curious and would lessen the intensity of grooming efforts when the child appeared hesitant [19].

Within online grooming, authors have also identified several characteristics related to the pacing and distribution of grooming. Through examining the word categories present within offender chats, Black et al. [6] confirmed the findings of O'Connell [4] who determined stages are non-linear. Additionally, Black et. al [6] found within the first 20% of a chat an offender will go through multiple stages of the grooming process including potentially friendship forming, risk assessment, and sexual conversation. Kloess et al [13] described the pacing of offender conversations as varying. Some offenders formed a deep relationship with the minors and incorporated sexual topics slowly. Other offenders chose to not develop any relationship with the minor and instead engaged in sexual conversation with the minor almost immediately [13]. Similar to [13], Winters, Kaylor, and Jeglic [20] found sexual conversation was often mentioned early in conversation as a means to assess interest in sexual activity. Offenders within the study also attempted to arrange meetings within a short period of time [20].

Overall, qualitative research shows the distribution of grooming themes and stages within a chat conversation varies [6], [13], [20]. Additionally, the intensity of sexual conversation and the level of risk of physical meeting varies throughout the conversation and does not instantly change but rather escalates and de-escalates depending upon the situation [4], [13], [20].

Within the computational analysis of offender chats, a large portion of research has focused on automatic triage of offenders within chats [8], [9], [11], [21]. While features for identification of offender conversations vary, one strategy is to operationalize themes and stages of the grooming process through natural language processing tools.

Cano, Fernandez, and Alani [22] use various behavioral and lexical features to classify lines of chats into the stages of Luring Communication Theory, which maps online grooming themes to a communication process centered around deceptive trust [5]. The three stages included grooming, approach, and trust development[22]. The features used included sentiment, bag of words, readability, lexical category features, and chat patterns. The authors found combining features together resulted in higher performance when detecting both the grooming phase and the approach phase. Additionally, the authors found sentiment features were not good at discriminating between stages. Lexical category features were found to improve classification but on their own were not good predictors of the phases[22].

A similar study was performed by Michalopoulos and Mavridis [23] in which the authors also attempted to classify the elements of grooming based on a set of features. In this study, the stages being categorized were sexual affair, gaining access, and deceptive relationship. The authors used TF-IDF and other document classification methods [23].

Finally, McGhee et al. [21] compared their rules-based approach to identifying the pursuit of personal information, the grooming process, and the approach of the predator to the victim to a machine learning approach which used a series of features including various categories of words and various groups of parts of speech. Some of the categories of words included approach nouns, activity nouns, and family nouns. The categories were similar to the categories which are found in the lexical category dictionary in the Linguistic Word Count (LIWC) tool [24]. The authors found their previous work which used a rule-based approach was superior to the use of the machine learning methods of decision trees and instance based learning [21]. However, the accuracy of the rule-based system only reaches 68% which leaves room for improvement[21].

The majority of the computational studies rely on surface level features. However, multiple studies have found combining such features, or using features which are representative

of the grooming process is a more successful approach[11], [22]. Identifying only features such as n-grams, part of speech, and word categories does not provide an accurate picture of grooming because the strategies within an offender conversation (1) do not occur in a particular order, (2) are gradually shifted between, and (3) are repeated in various combinations throughout the chat based on the input of the minor [4], [6], [17]. As a result, annotating grooming characteristics for computational use is a crucial task. However, manual annotation of grooming characteristics is difficult for the same reasons. The ebb and flow of themes causes issues for annotators in terms of both complexity and code span [14], [15]. Issues with complexity and code span result in lower inter-annotator reliability, which affects the corpus's usability and validity [14], [15]. For such a task, a fuzzy representation of annotations is preferable, as lines may be progressive transitions between or within stages of the grooming process.

III. METHODOLOGY

A. Building the Corpus

To build our corpus, we choose eight chats from Perverted Justice (PJ), comprising a total of 13,648 individual messages, with 16 unique users. PJ is a vigilante organization which seeks to identify sexual solicitors of minors online. The participants act as child decoys and talk to potential solicitors as if they are minors. If the conversation progresses to a level in which an individual is considered dangerous and/or breaks the law, the organization contacts and works with law enforcement to potentially capture the individual. The chat conversations occurring between convicted offenders and decoys are posted to the Perverted Justice website and are the primary source of chat conversations for individuals studying online solicitor communication [6], [25], [26].

As we are interested in improving the triage process of chats for law enforcement, we annotate the eight offender-decoy chats for risk. The literature leads us to believe fuzzy representations of annotations are preferable. However before fuzzy sets can be built, we need to start with a set of crisp labels. We labeled each line of the eight chat conversations as being low, medium, or high risk.

1) Low Risk Lines: These were defined as those lines which could be considered typical of a non-sexual chat between two individuals. In terms of grooming, this corresponded to the friendship forming stage, relationship forming stage, and risk assessments which did not obviously appear to be related to sex or deception (E.g. requesting an image of a child that was non-sexual).

Fig 1. is an example of a set of low risk lines occurring at the beginning of one of the conversations;

Solicitor: hey
Decoy: hey. ur in jasper?
Solicitor: yes
Decoy: kool wats u doin
Solicitor: nothing
Solicitor: i'm just laying in bed

Fig. 1: A snippet of low-risk messages

The excerpt was considered to be low risk because the solicitor engaged in small talk typical of two individuals meeting in a chat room.

2) Medium risk lines: These are lines in which explicit affection, compliments about the decoys appearance or body parts which are implicitly but not explicitly sexual, secrecy, guilt, or the exclusivity stage are present. The medium level of risk maps to language around an isolated, covert bond being formed between the offender and the decoy. The following lines were labeled as medium risk within one of the chats (Fig. 2):

Solicitor: look at you just a lil angel
Solicitor: lol
Decoy: thanks :p
Solicitor: i think my fav is you in the green tights
Solicitor: well i like them all actually
Decoy: thanks yeah it shows the most of me
Solicitor: yeah a lil bit of your side
Solicitor: lol
Decoy: lol yeah i bet you like that ¿:)
Solicitor: yeah i do

Fig. 2: A snippet of medium-risk messages

In this section of the chat, the offender compliments the decoy by calling her a *lil angel* and makes comments about the clothing of the decoy in the image she had sent. The decoy implies the solicitor likes the image for sexual reasons, but does not directly state this. The solicitor responds in the affirmative, but does not explicitly sexualize the conversation.

3) High risk lines: These are lines in which the sexual stage or discussion of meeting and explicit risk assessment of meeting occur. In the previous excerpt, the solicitor/decoy was headed in a sexual direction but was not explicit. In the following lines, the offender encourages the minor to meet physically (Fig 3):

Solicitor: i'm soo bored ..i'm coming to get u
Solicitor: jk
Solicitor: ouch ..good move
Decoy: ohhh ur jk?lol
Solicitor: unless u want me to ;)

Fig. 3: A snippet of high-risk messages

The above excerpt shows an offender introducing the concept of meeting by framing it as a joke. As the conversation continued, the offender began asking more and more hypotheticals related to whether or not the child would like to meet and what the meeting would entail. This progression is common within offender chats. While the initial coding of the corpus is crisp, the actual transitions between risk levels is not. For instance, in Fig. 4, the offender and victim start casually talking and the offender incorporates affectionate talk which is consistent with the exclusivity stage.

Based on the conversations that were used in this work, in addition to previous literature, we discovered that three lines before and after the beginning of each definitive risk

Solicitor: hey
Decoy: hey
Solicitor: hey what are you up to
Solicitor: you there
Solicitor: really where you been at
Decoy: just not here much. what r u up to?
Solicitor: missing you
Decoy: u really missed me?
Solicitor: yes i try to call your number but you never answer

Fig. 4: A snippet of a risk transition

level chunk would approximately contribute in the transition between risk levels [14], [15]. Thus, we start building our risk level fuzzy sets starting from the fourth line before and after the beginning of each crisp label for the risk. In the next section, we discuss the methodology for creating fuzzy sets from the crisp labels we initially created.

B. Building Fuzzy Sets from Crisp Labels

The fuzzy set for each line comprises of the line's membership degree to each class - low, medium or high. To build our fuzzy set from the imprecise crisp annotations, we use a trapezoidal function, as indicated by Eq (1):

$$\mu_C(l) = \begin{cases} \frac{l-a}{4} & \text{if } a-4 \leq l < a \\ 1 & \text{if } a \leq l \leq b \\ \frac{b+4-l}{4} & \text{if } b < l \leq b+4 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where l indicates maximum membership of the line to the given class C, a and b indicate the start and ending points of the crisp set. For the three classes C_{low}, C_{medium}, and C_{high}, the fuzzy set describing the membership of risk for line l would be the following vector: $\mu(l) = [\mu_{low}(l), \mu_{medium}(l), \mu_{high}(l)]^\top$ Hence, models used in the task of measuring risk would estimate the membership of each class rather than classify the line into a singular category.

IV. MODELS FOR FUZZY RISK ESTIMATION

In this section, we propose simple models that have been shown to be competitive in regular text classification tasks. These models are modified variants of previously successful models that have shown to be simple and difficult to beat baselines for sentence classification. Specifically, we use two models:

A. Model Description

1) Deep Averaging Network (DAN): As described in Iyyer et al. [27],], this model takes the embeddings of each word in the input and averages them to produce a sentence representation, which is then passed as input into a Multi-Layer Feed-Forward Network with a Softmax layer at the end to classify the input. This model has shown to be a strong baseline for many classification tasks [27].

2) Multi-Channel Convolution Neural Networks (CNN): Using CNNs for sentence classification is a technique first proposed by Kim [28]. In this model, each word in the sentence is initialized by a word-based pretrained representation as well as a random vector, which comprise of the two channels, and multiple 1D convolution filters are used to extract signals from various n-grams in the sentence, which are passed through a max-pooling layer to produce a fixed length sentence representation which is finally used for classification using a standard feed-forward layer.

3) Modifications to the above models: In our case, we use fasttext vectors [29] to represent words in DAN as well as in initializing the CNN model. Additionally, for words that do not have a vector in the pretrained embedding matrix, we use fasttexts approximation by summing the unknown words sub-word vectors (3-6 character n-grams) to represent the word. Since we are trying to estimate the entire fuzzy set instead of a single class, we use the sigmoid function in the last layer of both these models, and return a 3-dimensional vector comprising of the membership values of the input for each risk-level.

B. Loss Function

To train our networks, we use the L1 Loss function to compare our output vector with the ground truth fuzzy set. The L1 loss function computes the sum of absolute-differences between two vectors, it is also referred to as Least Absolute Deviations, and is robust when it encounters outliers. Our models are trained by jointly minimizing the following function, for all classes (low, medium, high):

$$L = \sum_{i \in C} |\mu_i(l) - \hat{y}_i| \quad (2)$$

Where $\mu_i(l)$ and $\hat{y}_i$ are the ground truth and the estimated value of the membership for class i of the given line, where $i \in C = \{low, medium, high\}$

C. Evaluation

The ground truth for risk of a given line is a fuzzy set with membership values of the line for three classes, and our proposed baselines produce a 3-dimensional vector by training using the input chat message and ground truth. Since these are not singular, precise values as observed in binary/multinomial classification, metrics such as accuracy, precision, F1 scores are not suitable to evaluate our results. Thus, we use a fuzzy metric to calculate how similar the output of the neural network is to the truth fuzzy set of the line. In this case, we use a Fuzzy Jaccard Similarity metric [30], where fuzzified versions of the union, intersections and cardinality are used. Mathematically, the Jaccard Similarity between two sets, A and B, is given by:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

In fuzzy operations [16] between two fuzzy sets A and B with membership μ_i, denoting the set's membership degree for the i^{th} class, their intersection, $A \cap B$, is given by

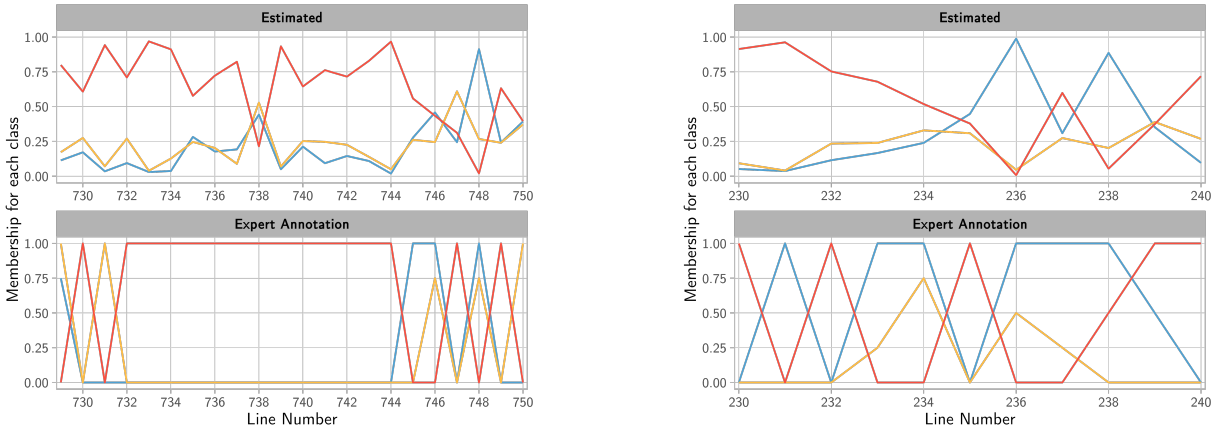

Fig. 5: Illustration of the truth (bottom) membership values, and values estimated by the CNN model (top) on randomly selected chunks in the test set. **Low:** Blue, **Medium:** Yellow, **High:** Red

$\min\{\mu_i(A), \mu_i(B)\}_{i \in C}$, their union, $A \cup B$, is given by $\max\{\mu_i(A), \mu_i(B)\}_{i \in C}$. Finally, the cardinality of a fuzzy set, say Q, denoted by $|Q|$, is given by $\sum_1^C \mu_i(Q)$. Thus, the Fuzzy Jaccard Similarity between fuzzy sets A and B, is computed as follows:

$$J_{Fuzzy}(A, B) = \frac{\sum\limits_{i \in C} \min\left\{\mu_i(A), \mu_i(B)\right\}_{i \in C}}{\sum\limits_{i \in C} \max\left\{\mu_i(A), \mu_i(B)\right\}_{i \in C}} \quad (4)$$

For our purposes, we calculate this metric for each line, and report the average fuzzy-Jaccard similarity over the entire set (training. validation or test). It is bounded by 0 and 1, i.e., the closer the output produced by the model to the true-fuzzy set for a given line, the greater is the value of this metric. Thus, it serves as a good evaluation metric to compare different models over a large collection of lines, as is in our case.

D. Experiments and Results

We split the final corpus into train, validation, and test sets. Out of the 8 different conversations, we randomly select two whole conversations to be our validation and test sets respectively. We select conversations instead of chunks of lines or randomly sampled lines because during our analysis, we can explore and probe the model based on an entire meaningful unit, in this case a conversation, rather than assess how it performed on isolated lines. Table I describes the number of lines in each of the sets.

TABLE I: NUMBER OF LINES IN EACH SET

Set	Conversations	Size(lines)
Train	6	11900
Valid.	1	977
Test	1	771

We train our DAN and CNN models with a dropout of 0.5 after the embedding layer, a learning rate of 1×10^{-4}, and using the Adam Optimizer. The DAN model was trained for 1000 epochs with a minibatch size of 50, while the CNN model was trained for 100 epochs with a minibatch size of 32. The best model was chosen based on the highest average Fuzzy-Jaccard Similarity on the validation set. The results of both models are shown in Table II.

TABLE II: RESULTS OF MODELS ON TEST SET

Model	Epochs	J_{fuzzy}
CNN	100	**0.455**
DAN	1000	0.380

Figure 5 shows excerpts from two of the chats in the test set. While the ground truth and the model's prediction look somewhat different, the model makes sense because the various grooming stages are not independent [6], [4]. Elements of each of the grooming stages feeds into the next. For instance, the exclusivity stage is focused around trust and isolation which is used to transition the child into the sexual stage [4]. So, it is feasible for a high risk line to contain elements of low and medium risk because the high risk sexual or meeting information build upon the low and medium risk information.

The estimated labels in Figure 5B demonstrate the progressive decrease and increase of sexual topics throughout the chat. As the high risk level decreases, the low risk and medium risk related to the non-sexual aspects of the grooming level increase. This is consistent with previous research which indicates grooming is non-linear and gradually progresses between stages [4], [6], [17]. The estimated labels in Figure 5A are also consistent with the ebb and flow of the grooming process and also accurately capture the sexual nature of the chat from lines 730 745 as well as the variability of the last 5 lines.

V. CONCLUSIONS

In this work, our aim was to estimate risk of sexual predation within online chat conversations. We deviated from the classical notion of crisp sets and instead use fuzzy membership functions to quantify the amount of risk present

in each chat message. In our framework, a given chat message can belong to low, medium and high risk with varying membership degrees. We experimented on eight online conversations between predators and decoys by using a fuzzy membership function to label the amount of risk within each line, which were used in two neural network models that were tested on a new conversation. While the models achieved moderate values of the fuzzy-jaccard similarity, the patterns of the various risks as produced from the model on a new conversation are consistent with existing literature on grooming processes. Leveraging contextual cues and the relationships between the various message snippets can help these models perform better.

ACKNOWLEDGEMENTS

This research was partially supported by Purdue Research Foundation.

REFERENCES

[1] "Online Enticement of Children: An In-Depth Analysis of CyberTipLine Reports," in, National Center for Missing and Exploited Children, 2017.

[2] H. C. Whittle, C. Hamilton-Giachritsis, and A. R. Beech, "Victimsfffdfffdfffd voices: The impact of online grooming and sexual abuse," *Universal Journal of Psychology*, vol. 1, no. 2, pp. 59–71, 2013.

[3] M. M. Chiu, K. C. Seigfried-Spellar, and T. R. Ringenberg, "Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: Statistical discourse analysis of self-disclosure and emotion words," *Child abuse & neglect*, vol. 81, pp. 128–138, 2018.

[4] R. OĆonnell, "A typology of child cybersexploitation and online grooming practices," *Preston, UK: University of Central Lancashire*, 2003.

[5] L. N. Olson, J. L. Daggs, B. L. Ellevold, and T. Rogers, "Entrapping the innocent: Toward a theory of child sexual predators' luring communication," *Communication Theory*, vol. 17, no. 3, pp. 231–251, 2007.

[6] P. J. Black, M. Wollis, M. Woodworth, and J. T. Hancock, "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*, vol. 44, pp. 140–149, 2015.

[7] K. M. Babchishin, R. K. Hanson, and H. VanZuylen, "Online child pornography offenders are different: A meta-analysis of the characteristics of online and offline sex offenders against children," *Archives of sexual behavior*, vol. 44, no. 1, pp. 45–66, 2015.

[8] J. Parapar, D. E. Losada, and A. Barreiro, "A learning-based approach for the identification of sexual predators in chat logs.," in *CLEF)*, vol. 1178, 2012.

[9] M. Ebrahimi, C. Y. Suen, and O. Ormandjieva, "Detecting predatory conversations in social media by deep convolutional neural networks," *Digital Investigation*, vol. 18, pp. 33–49, 2016.

[10] C. Barber and S. Bettez, "Deconstructing the online grooming of youth: Toward improved information systems for detection of online sexual predators," 2014.

[11] D. Bogdanova, P. Rosso, and T. Solorio, "Exploring high-level features for detecting cyberpedophilia," *Computer speech & language*, vol. 28, no. 1, pp. 108–120, 2014.

[12] P. S. Bayerl and K. I. Paul, "What determines inter-coder agreement in manual annotations? a meta-analytic investigation," *Computational Linguistics*, vol. 37, no. 4, pp. 699–725, 2011.

[13] J. A. Kloess, S. Seymour-Smith, C. E. Hamilton-Giachritsis, M. L. Long, D. Shipley, and A. R. Beech, "A qualitative analysis of offenders' modus operandi in sexually exploitative interactions with children online," *Sexual Abuse*, vol. 29, no. 6, pp. 563–591, 2017.

[14] P. Kingsbury, M. Palmer, and M. Marcus, "Adding semantic annotation to the penn treebank," in *Proceedings of the human language technology conference*, Citeseer, 2002, pp. 252–256.

[15] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, *et al.*, "Concept annotation in the craft corpus," *BMC bioinformatics*, vol. 13, no. 1, p. 161, 2012.

[16] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[17] A. Gupta, P. Kumaraguru, and A. Sureka, "Characterizing pedophile conversations on the internet using online grooming," *arXiv preprint arXiv:1208.4324*, 2012.

[18] J. J. Peters, "Children who are victims of sexual assault and the psychology of offenders," *American Journal of Psychotherapy*, vol. 30, no. 3, pp. 398–421, 1976.

[19] R. A. Lang and R. R. Frenzel, "How sex offenders lure children," *Annals of Sex Research*, vol. 1, no. 2, pp. 303–317, 1988.

[20] G. M. Winters, L. E. Kaylor, and E. L. Jeglic, "Sexual offenders contacting children online: An examination of transcripts of sexual grooming," *Journal of sexual aggression*, vol. 23, no. 1, pp. 62–76, 2017.

[21] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, "Learning to identify internet sexual predation," *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 103–122, 2011.

[22] A. Cano Basave, M. Fernández, and H. Alani, "Detecting child grooming behaviour patterns on social media," 2014.

[23] D. Michalopoulos and I. Mavridis, "Utilizing document classification for grooming attack recognition," in *2011 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, 2011, pp. 864–869.

[24] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[25] K. Guice, *Predators, decoys, and teens: A corpus analysis of online language*. Hofstra University, 2016.

[26] M. Drouin, R. L. Boyd, J. T. Hancock, and A. James, "Linguistic analysis of chat transcripts from child predator undercover sex stings," *The Journal of Forensic Psychiatry & Psychology*, vol. 28, no. 4, pp. 437–457, 2017.

[27] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 1681–1691.

[28] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[30] V. Zhelezniak, A. Savkov, A. Shen, F. Moramarco, J. Flann, and N. Y. Hammerla, "Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors," in *International Conference on Learning Representations*, 2019.